An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity

Cartis, Coralia; Gould, Nick; Toint, Philippe

Published in:
IMA Journal of Numerical Analysis

DOI:
10.1093/imanum/drr035

Publication date:
2012

Document Version
Early version, also known as pre-print

An adaptive cubic regularisation algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity

by C. Cartis[1], N. I. M. Gould[2] and Ph. L. Toint[3]

Report 08/05(revised) 18 February 2009

[1] Department of Mathematics,
University of Edinburgh,
Edinburgh, United Kingdom.
Email: coralia.cartis@ed.ac.uk

[2] Computational Science and Engineering Department
Rutherford Appleton Laboratory
Chilton OX11 0QX, United Kingdom.
Email: nick.gould@stfc.ac.uk

[3] Department of Mathematics,
FUNDP-University of Namur,
61, rue de Bruxelles, B-5000 Namur, Belgium.
Email: philippe.toint@fundp.ac.be

An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity

C. Cartis, N. I. M. Gould and Ph. L. Toint

18 February 2009

Abstract

The adaptive cubic overestimation algorithm described in Cartis, Gould and Toint (2007) is adapted to the problem of minimizing a nonlinear, possibly nonconvex, smooth objective function over a convex domain. Convergence to first-order critical points is shown under standard assumptions, but without any Lipschitz continuity requirement on the objective's Hessian. A worst-case complexity analysis in terms of evaluations of the problem's function and derivatives is also presented for the Lipschitz continuous case and for a variant of the resulting algorithm. This analysis extends the best known bound for general unconstrained problems to nonlinear problems with convex constraints.

Keywords: Nonlinear optimization, convex constraints, cubic regularisation, numerical algorithms, global convergence, worst-case complexity.

1 Introduction

Adaptive cubic regularisation has recently returned to the forefront of smooth nonlinear optimization as an alternative to more standard globalization techniques for nonlinear unconstrained optimization. Methods of this type, initiated by Griewank (1981), Nesterov and Polyak (2006) and Weiser, Deuflhard and Erdmann (2007), have been consolidated into a practical and successful algorithm by Cartis, Gould and Toint (2007). They are based on the observation that a third-order model can be constructed which is an overestimate of the objective function when the latter has Lipschitz continuous Hessians and a model parameter is chosen large enough. These adaptive overestimation methods are not only globally convergent to first- and second-order critical points, but also enjoy good worse-case complexity bounds. Furthermore, numerical results presented in Cartis et al. (2007) suggest that it might be one of the most efficient numerical minimization methods to date.

Extending the approach to more general optimization problems is therefore attractive, as one may hope that some of the qualities of the unconstrained methods can be transferred to a broader framework. Nesterov (2006) has considered the extension of his cubic regularisation method to problems with smooth convex objective function and convex constraints. In this paper, we consider that of the adaptive cubic overestimation method to the case where minimization is subject to convex constraints, but the smooth objective function is no longer assumed to be convex. The new algorithm is strongly inspired by the unconstrained adaptive cubic overestimation method and by the trust-region projection methods for the same problem, which are fully described in Chapter 12 of Conn, Gould and Toint (2000). In particular, it makes significant use of the specialized criticality measure developed by Conn, Gould, Sartenaer and Toint (1993) for this context. Remarkably, the desirable iteration complexity of the cubic regularisation method for unconstrained nonlinear problem extends to the case where convex constraints are present. Because the number of objective function/gradient evaluations is directly dependent on the number and type of the iterations, one therefore deduces a worst-case bound of the number of these evaluations.

The paper is organized as follows. Section 2 describes the problem more formally as well as the new algorithm, while Section 3 presents the associated convergence theory (to first-order critical points). We then discuss a worst-case function-evaluation complexity result for a variant of the new algorithm in Section 4. Some conclusions are finally presented in Section 5.

2 The new algorithm

We consider the numerical solution of the constrained nonlinear optimization problem

$$\min_{x\in\mathcal{F}} f(x), \tag{2.1}$$

where we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, possibly nonconvex, and bounded below by the constant $f_{\rm low}$ on the closed convex non-empty feasible domain $\mathcal{F} \subseteq \mathbb{R}^n$.

Our algorithm for solving this problem follows the broad lines of the projection-based trust-region algorithm of Chapter 12 in Conn et al. (2000) with adaptations necessary to replace the trust-region globalization mechanism by a cubic regularisation of the type analysed in Cartis et al. (2007). At an iterate x_k within the feasible region $\mathcal{F}$, a cubic model of the form

$$m_k(x_k+s) = f(x_k) + \langle g_k, s_k\rangle + \tfrac{1}{2}\langle s_k, B_k s_k\rangle + \tfrac{1}{3}\sigma_k\|s_k\|^3 \tag{2.2}$$

is defined, where $\langle\cdot,\cdot\rangle$ denotes the Eclidean inner product, where $g_k \stackrel{\rm def}{=} \nabla_x f(x_k)$, where B_k is a symmetric matrix hopefully approximating $\nabla_{xx} f(x_k)$, where σ_k is a non-negative regularisation parameter, and where $\|\cdot\|$ stands for the Euclidean norm. The step s_k from x_k is then defined in two stages. The first stage is to compute a *generalized Cauchy point* $x_k^{\rm GC}$ such that $x_k^{\rm GC}$ approximately minimizes the model (2.2) along the Cauchy arc defined by the projection onto $\mathcal{F}$ of the negative gradient path, that is

$$\{\, x \in X \mid x = P_{\mathcal{F}}[x_k - t g_k] \,\},$$

where we define $P_{\mathcal{F}}$ to be the (unique) orthogonal projector onto $\mathcal{F}$. The approximate minimization is carried out using a generalized Goldstein-like linesearch on the arc, as explained in Section 12.1 of Conn et al. (2000). In practice, $x_k^{\rm GC} = x_k + s_k^{\rm GC}$ is determined such that

$$x_k^{\rm GC} = P_{\mathcal{F}}[x_k - t_k^{\rm GC} g_k] \ \text{ for some } \ t_k^{\rm GC} > 0, \tag{2.3}$$

and

$$m_k(x_k^{\rm GC}) \le f(x_k) + \kappa_{\rm ubs}\langle g_k, s_k^{\rm GC}\rangle \tag{2.4}$$

and either

$$m_k(x_k^{\rm GC}) \ge f(x_k) + \kappa_{\rm lbs}\langle g_k, s_k^{\rm GC}\rangle \tag{2.5}$$

or

$$\|P_{T(x_k^{\rm GC})}[-g_k]\| \le \kappa_{\rm epp}|\langle g_k, s_k^{\rm GC}\rangle|, \tag{2.6}$$

where the three constants satisfy

$$0 < \kappa_{\rm ubs} < \kappa_{\rm lbs} < 1, \ \text{ and } \ \kappa_{\rm epp} \in (0, \tfrac{1}{2}). \tag{2.7}$$

and where $T(x)$ is the tangent cone to $\mathcal{F}$ at x. The conditions (2.4) and (2.5) are the familiar Goldstein linesearch conditions adapted to our search along the Cauchy arc, while (2.6) is there to handle the case where this arc ends before condition (2.5) is ever satisfied. Once the generalized Cauchy point $x_k^{\rm GC}$ is computed (which can be done by a suitable search on $t_k^{\rm GC} > 0$ inspired by Algorithm 12.2.2 of Conn et al. (2000) and discussed below), any step s_k such that

$$x_k^+ \stackrel{\rm def}{=} x_k + s_k \in \mathcal{F}$$

and such that the model value at x_k^+ is below that obtained at $x_k^{\rm GC}$ is acceptable.

Given the step s_k, the trial point x_k^+ is known and the value of the objective function at this point computed. If the ratio

$$\rho_k = \frac{f(x_k) - f(x_k^+)}{f(x_k) - m_k(x_k^+)}. \tag{2.8}$$

of the achieved reduction in the objective function compared to the predicted model reduction is larger than some constant $\eta_1 > 0$, then the trial point is accepted as the next iterate and the regularisation parameter σ_k essentially unchanged or increased, while the trial point is rejected and σ_k increased if

$\rho_k < \eta_1$. Fortunately, the undesirable situation where the trial point is rejected cannot persist since σ_k eventually becomes larger than some local Lipschitz constant associated with the Hessian of the objective function (assuming it exists), which in turn guarantees that $\rho_k \geq 1$, as shown in Griewank (1981), Nesterov and Polyak (2006) or Cartis et al. (2007).

We now state our Adaptive Cubic Regularisation for Convex constraints (ACURC).

Algorithm 2.1: Adaptive Cubic Regularisation for Convex Constraints (ACURC)

Step 0: Initialization. An initial point $x_0 \in \mathcal{F}$ and an initial regularisation parameter σ_0 are given. Compute $f(x_0)$ and set $k = 0$.

Step 1: Determination of the generalized Cauchy point. If x_k is first-order critical, terminate the algorithm. Otherwise perform the following iteration.

Step 1.0: Initialization. Define the model (2.2), choose $t_0 > 0$ and set $t_{\min} = 0$, $t_{\max} = \infty$ and $j = 0$.

Step 1.1: Compute a point on the projected-gradient path. Set $x_{k,j} = P_{\mathcal{F}}[x_k - t_j g_k]$ and evaluate $m_k(x_{k,j})$.

Step 1.2: Check for the stopping conditions. If (2.4) is violated, then set $t_{\max} = t_j$ and go to Step 1.3. Otherwise, if (2.5) and (2.6) are violated, set $t_{\min} = t_j$ and go to Step 1.3. Otherwise, set $x_k^{\rm GC} = x_{k,j}$ and go to Step 2.

Step 1.3: Find a new value of the arc parameter. If $t_{\max} = \infty$, set $t_{j+1} = 2t_j$. Otherwise, set $t_{j+1} = \frac{1}{2}(t_{\min} + t_{\max})$. Increment j by one and go to Step 1.2.

Step 2: Step calculation. Compute a step s_k and a trial point $x_k^+ \stackrel{\rm def}{=} x_k + s_k \in \mathcal{F}$ such that

$$m_k(x_k^+) \leq m_k(x_k^{\rm GC}). \tag{2.9}$$

Step 3: Acceptance of the trial point. Compute $f(x_k^+)$ and the ratio (2.8). If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$.

Step 4: Regularisation parameter update. Set

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases}$$

Increment k by one and go to Step 1.

As in Cartis et al. (2007), the constants η_1, η_2, γ_1, and γ_2 are given and satisfy the conditions

$$0 < \eta_1 \leq \eta_2 < 1 \;\text{ and }\; 1 < \gamma_1 \leq \gamma_2. \tag{2.10}$$

As for trust-region algorithms, we say that iteration k is successful whenever $\rho_k \geq \eta_1$ (and thus $x_{k+1} = x_k^+$), and very successful whenever $\rho_k \geq \eta_2$, in which case, additionally, $\sigma_{k+1} \leq \sigma_k$. We denote the index set of all successful iterations by $\mathcal{S}$.

As mentioned above, our technique for computing the generalized Cauchy point is inspired from the Goldstein linesearch scheme, but it is most likely that techniques based on Armijo-like backtracking (see Sartenaer, 1993) or on successive exploration of the active faces of $\mathcal{F}$ along the Cauchy arc (see Conn, Gould and Toint, 1988) are also possible, the latter being practical when $\mathcal{F}$ is a polyhedron.

3 Global convergence to first-order critical points

We now consider the global convergence properties of Algorithm ACURC and show in this section that all the limit points of the sequence of its iterates must be first-order critical points for the problem (2.1). Our analysis will be based on the first-order criticality measure at $x \in \mathcal{F}$ given by

$$\chi(x) \stackrel{\rm def}{=} \left| \min_{x+d\in\mathcal{F},\|d\|\le 1} \langle \nabla_x f(x), d\rangle \right|, \tag{3.1}$$

(see Conn et al., 1993) and define $\chi_k \stackrel{\rm def}{=} \chi(x_k)$. For our analysis, we consider the following assumptions.

AS1: The feasible set $\mathcal{F}$ is closed, convex and non-empty.

AS2: The function f is twice continuously differentiable on an open set $\hat{\mathcal{F}}$ containing $\mathcal{F}$.

AS3: The function f is bounded below by $f_{\rm low}$ on $\mathcal{F}$.

AS4: There exists constant $\kappa_{\rm H} > 1$ and $\kappa_{\rm B} > 1$ such that

$$\|\nabla_{xx} f(x)\| \le \kappa_{\rm H} \text{ for all } x \in \mathcal{F}, \text{ and } \|B_k\| \le \kappa_{\rm B} \text{ for all } k \ge 0. \tag{3.2}$$

Our first result investigate the properties of the projected gradient path and variants of the criticality measure (3.1).

Lemma 3.1 *Suppose that AS1 and AS2 hold. For $x \in \mathcal{F}$ and $t > 0$, let*

$$x(t) \stackrel{\rm def}{=} P[x - t\nabla_x f(x)] \text{ and } \theta(x,t) \stackrel{\rm def}{=} \|x(t) - x\|, \tag{3.3}$$

while, for $x \in \mathcal{F}$ and $\theta > 0$,

$$\chi(x,\theta) \stackrel{\rm def}{=} \left| \min_{x+d\in\mathcal{F},\|d\|\le \theta} \langle \nabla_x f(x), d\rangle \right|, \tag{3.4}$$

and

$$\pi(x,\theta) \stackrel{\rm def}{=} \frac{\chi(x,\theta)}{\theta}, \tag{3.5}$$

and

$$\pi_k^{GC} \stackrel{\rm def}{=} \pi(x_k, \|s_k^{GC}\|) \text{ and } \pi_k^{+} \stackrel{\rm def}{=} \pi(x_k, \|s_k\|), \tag{3.6}$$

where $s_k^{GC} \stackrel{\rm def}{=} x_k^{GC} - x_k$. Then $\theta(x,t)$, $\chi(x,\theta)$ and $\pi(x,\theta)$ are continuous with respect to their two arguments, $\theta(x,t)$ is non-decreasing as a function of t, $\chi(x,\theta)$ is non-decreasing with θ and $\pi(x,\theta)$ is non-increasing with θ. In particular, if $\|s_k^{GC}\| \ge 1$, then

$$\chi(x_k, \|s_k^{GC}\|) \ge \chi_k \ge \pi_k^{GC} \tag{3.7}$$

while if $\|s_k^{GC}\| \le 1$, then

$$\pi_k^{GC} \ge \chi_k \ge \chi(x_k, \|s_k^{GC}\|). \tag{3.8}$$

Similarly, if $\|s_k\| \ge 1$, then

$$\chi(x_k, \|s_k\|) \ge \chi_k \ge \pi_k^{+} \tag{3.9}$$

while if $\|s_k\| \le 1$, then

$$\pi_k^{+} \ge \chi_k \ge \chi(x_k, \|s_k\|). \tag{3.10}$$

Moreover

$$\chi_k \le \chi(x_k, \|s_k^{GC}\|) + 2\|P_{T(x_k^{GC})}[-g_k]\|, \tag{3.11}$$

$$-\langle g_k, s_k^{GC}\rangle = \chi(x_k, \|s_k^{GC}\|) \ge 0 \tag{3.12}$$

and

$$\theta(x,t) \ge t\,\|P_{T(x(t))}[-\nabla_x f(x)]\| \tag{3.13}$$

for all $t > 0$.

Proof. These results only depend on the geometry of the projected gradient path and (except for (3.13)) immediately follow from Theorems 12.1.3 (page 446), 12.1.4 (page 447) and 12.1.5 (page 448) in Conn et al. (2000) and the identity $\chi_k = \chi(x_k, 1)$. In particular, (3.11) results from (3.7) if $\|s_k^{\rm GC}\| \geq 1$, and from (3.12) and Th. 12.1.5 (iii) with $\theta = 1$ and $d = s_k^{\rm GC}$ if $\|s_k^{\rm GC}\| < 1$. We therefore only need to prove (3.13). We first note that, if $u(x,t) = x(t) - x$, then $\theta(x,t) = \|u(x,t)\|$ and, denoting the right directional derivative by d/dt_+, we see that

$$\frac{d\theta}{dt_+}(x,t) = \frac{\langle \frac{du(x,t)}{dt_+}, u(x,t)\rangle}{\|u(x,t)\|} = \frac{\langle P_{T(x(t))}[-\nabla_x f(x)], u(x,t)\rangle}{\theta(t)} \tag{3.14}$$

where we used Proposition 5.3.5 (page 141) of Hiriart-Urruty and Lemaréchal (1993) to deduce the second equality. Moreover

$$u(x,t) = -t\nabla_x f(x) - [x - t\nabla_x f(x) - x(t)] \stackrel{\rm def}{=} -t\nabla_x f(x) - z(x,t) \tag{3.15}$$

and because of the definition of $x(t)$, $z(x,t)$ must belong to $N(x(t))$, the normal cone to $\mathcal{F}$ at $x(t)$. Thus, since this cone is the polar of $T(x(t))$, we deduce that

$$\langle P_{T(x(t))}[-\nabla_x f(x)], z(x,t)\rangle \leq 0. \tag{3.16}$$

We now obtain, successively using (3.14), (3.15) and (3.16), that

$$\begin{array}{rcl} \theta(t)\dfrac{d\theta}{dt_+}(t) & = & \langle P_{T(x(t))}[-\nabla_x f(x)], u(x,t)\rangle \\ & = & \langle P_{T(x(t))}[-\nabla_x f(x)], -t\nabla_x f(x) - z(x,t)\rangle \\ & = & t\,\langle -\nabla_x f(x), P_{T(x(t))}[-\nabla_x f(x)]\rangle - \langle P_{T(x(t))}[-\nabla_x f(x)], z(x,t)\rangle \\ & \geq & t\,\|P_{T(x(t))}[-\nabla_x f(x)]\|^2. \end{array} \tag{3.17}$$

But (3.14) and the Cauchy-Schwartz inequality also imply that

$$\frac{d\theta}{dt_+}(x,t) \leq \|P_{T(x(t))}[-\nabla_x f(x)]\|.$$

Combining this last bound with (3.17) finally yields (3.13) as desired. □

We complete our analysis of the criticality measures by considering the Lipschitz continuity of the measure $\chi(x)$. We start by proving an extension of Lemma 1 in Mangasarian and Rosen (1964).

Lemma 3.2 *Suppose that AS1 holds and define*

$$\phi(x) \stackrel{\rm def}{=} \min_{x+d\in\mathcal{F}, \|d\|\leq 1} \langle g, d\rangle$$

for $x \in \mathcal{F}$ and some vector $g \in I\!R^n$. Then $\phi(x)$ is a proper convex function on

$$\mathcal{F}_1 \stackrel{\rm def}{=} \{x \in I\!R^n \mid \|x - x_0\| \leq 1 \quad \textit{for some} \quad x_0 \in \mathcal{F}\} \supseteq \mathcal{F}. \tag{3.18}$$

Proof. The result is trivial if $g = 0$. Assume therefore that $g \neq 0$. We first note that the definition of $\mathcal{F}_1$ ensures that the feasible set of $\phi(x)$ is nonempty and therefore that the parametric minimization problem defining $\phi(x)$ is well-defined for any $x \in \mathcal{F}_1$. Moreover, the minimum is always attained because of the constraint $\|d\| \leq 1$, and so $-\infty < -\|g\| \leq \phi(x)$ for all $x \in \mathcal{F}_1$. Hence $\phi(x)$ is proper in $\mathcal{F}_1$. To show that $\phi(x)$ is convex, let $x_1, x_2 \in \mathcal{F}_1$, and let $d_1, d_2 \in I\!R^n$ be such that

$$\phi(x_1) = \langle g, d_1\rangle \quad \text{and} \quad \phi(x_2) = \langle g, d_2\rangle.$$

Also let $\lambda \in [0,1]$, $x_0 \stackrel{\rm def}{=} \lambda x_1 + (1-\lambda)x_2$ and $d_0 \stackrel{\rm def}{=} \lambda d_1 + (1-\lambda)d_2$. Let us show that d_0 is feasible for the $\phi(x_0)$ problem. Since d_1 and d_2 are feasible for the $\phi(x_1)$ and $\phi(x_2)$ problems, respectively, and since $\lambda \in [0,1]$, we have that $\|d_0\| \leq 1$. To show $x_0 + d_0 \in \mathcal{F}$; we have

$$x_0 + d_0 = \lambda(x_1 + d_1) + (1-\lambda)(x_2 + d_2) \in \lambda\mathcal{F} + (1-\lambda)\mathcal{F} \subseteq \mathcal{F},$$

where we used that $\mathcal{F}$ is convex to obtain the set inclusion. Thus d_0 is feasible for $\phi(x_0)$ and hence

$$\phi(x_0) \le \langle g, d_0\rangle = \lambda\langle g, d_1\rangle + (1-\lambda)\langle g, d_2\rangle = \lambda\phi(x_1) + (1-\lambda)\phi(x_2).$$

which proves that $\phi(x)$ is convex in $\mathcal{F}_1$. □

We are now is position to prove that the criticality measure $\chi(x)$ is Lipschitz continuous on bounded subsets of $\mathcal{F}$.

Theorem 3.3 *Supppose that AS1 and AS2 hold. Suppose also $\mathcal{F}_0$ is a bounded subset of $\mathcal{F}$ and that $\nabla_x f(x)$ is Lipschitz continuous on $\mathcal{F}_0$ with constant $\kappa_{\mathrm{L}g}$. Then there exists a constant $\kappa_{\mathrm{L}\chi} \ge 0$ such that*

$$|\chi(x) - \chi(y)| \le \kappa_{\mathrm{L}\chi}\|x-y\|$$

for all $x, y \in \mathcal{F}_0$.

Proof. We have from (3.1) that

$$\begin{aligned}\chi(x)-\chi(y) &= \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(y), d\rangle - \min_{x+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(x), d\rangle, &(3.19)\\ &= \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(y), d\rangle - \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(x), d\rangle\\ &\quad + \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(x), d\rangle - \min_{x+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(x), d\rangle. &(3.20)\end{aligned}$$

Note that the first two terms in (3.20) have the same feasible set but different objectives, while the last two have different feasible sets but the same objective. Consider the difference of the first two terms. Letting

$$\langle\nabla f(y), d_y\rangle = \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(y), d\rangle \quad\text{and}\quad \langle\nabla f(x), d_x\rangle = \min_{y+d\in\mathcal{F},\|d\|\le 1}\langle\nabla f(x), d\rangle,$$

the first difference in (3.20) becomes

$$\begin{aligned}\langle\nabla f(y), d_y\rangle - \langle\nabla f(x), d_x\rangle &= \langle\nabla f(y), d_y - d_x\rangle + \langle\nabla f(y)-\nabla f(x), d_x\rangle\\ &\le \langle\nabla f(y)-\nabla f(x), d_x\rangle\\ &\le \|\nabla f(y)-\nabla f(x)\|\cdot\|d_x\|\\ &\le \kappa_{\mathrm{L}g}\|x-y\|,\end{aligned} \qquad (3.21)$$

where to obtain the first inequality above, we used that, by definition of d_y and d_x, d_x is now feasible for the constraints of the problem of which d_y is the solution; the last inequality follows from the assumed Lipschitz continuity of ∇f and from the bound $\|d_x\| \le 1$.

Consider now the second difference in (3.20) (where we have the same objective but different feasible sets), and define

$$\mathcal{F}_{01} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid \|x - x_0\| \le 1 \quad\text{for some}\quad x_0 \in \mathcal{F}_0\}.$$

Note that our assumptions imply that $\mathcal{F}_{01}$ is a bounded subset of $\mathcal{F}_1$, where $\mathcal{F}_1$ is defined by (3.18). The proper convexity of $\phi(x)$ on $\mathcal{F}_1\cap\hat{\mathcal{F}}$ (ensured by Lemma 3.2 with $g = \nabla_x f(x)$) and Theorem 10.4 in Rockafellar (1970) then yield that $\phi(x)$ is Lipschitz continuous (with constant $\kappa_{\mathrm{L}\phi}$, say) on any subset of the relative interior of $\mathcal{F}_{01}\cap\hat{\mathcal{F}}$, in particular on $\mathcal{F}_0$. As a consequence, we obtain from (3.20) and (3.21) that

$$\chi(x) - \chi(y) \le (\kappa_{\mathrm{L}g} + \kappa_{\mathrm{L}\phi})\|x-y\|.$$

Since the role of x and y can be interchanged in the above argument, the conclusion of the theorem follows by setting $\kappa_{\mathrm{L}\chi} = \kappa_{\mathrm{L}g} + \kappa_{\mathrm{L}\phi}$. □

This theorem provides a generalization of a result already known for the special case where $\mathcal{F}$ is defined by simple bounds and the norm used in the definition of $\chi(x)$ is the infinity norm (see Lemma 4.1 in Gratton, Mouffe, Toint and Weber-Mendonça, 2008*a*).

We say that x_* is a first-order critical point for (2.1) if $\chi(x_*, 1) = 0$ (see Theorem 12.1.6 in Conn et al., 2000), and now prove a first crude upper bound on the length of any descent step.

Lemma 3.4 *Suppose that AS1 and AS2 hold, and that*

$$m_k(x_k + s) \le f(x_k). \tag{3.22}$$

Then

$$\|s\| \le \frac{3}{2\sigma_k}\left[\kappa_{\rm B} + \sqrt{\sigma_k \|g_k\|}\right]. \tag{3.23}$$

Proof. The definition (2.2) and (3.22) give that

$$\langle g_k, s\rangle + \tfrac{1}{2}\langle s, B_k s\rangle + \tfrac{1}{3}\sigma_k \|s\|^3 \le 0$$

and hence, using the Cauchy-Schwarz inequality and (3.2), that

$$0 \le \tfrac{1}{3}\sigma_k \|s\|^3 \le \|g_k\|\,\|s\| + \tfrac{1}{2}\kappa_{\rm B}\|s\|^2.$$

This in turn implies that

$$\|s\| \le \frac{\tfrac{1}{2}\kappa_{\rm B} + \sqrt{\tfrac{1}{4}\kappa_{\rm B}^2 + \tfrac{4}{6}\sigma_k\|g_k\|}}{\tfrac{2}{3}\sigma_k} = \frac{\kappa_{\rm B} + \sqrt{\tfrac{4}{6}\sigma_k\|g_k\|}}{\tfrac{2}{3}\sigma_k} \le \frac{3}{2\sigma_k}\left[\kappa_{\rm B} + \sqrt{\sigma_k\|g_k\|}\right].$$

□

Using this bound, we next verify that Step 1 of Algorithm ACURC is well-defined and delivers a suitable generalized Cauchy point.

Lemma 3.5 *Suppose that AS1 and AS2 hold. Then, for each k with $\chi_k > 0$, the loop between steps 1.1, 1.2 and 1.3 of Algorithm ACURC is finite and produces a generalized Cauchy point x_k^{GC} satisfying (2.4)-(2.6).*

Proof. Observe first that the generalized Cauchy point resulting from Step 1 must satisfy the conditions (2.4)-(2.6) if the loop on j internal to this step terminates finitely. Thus we only need to show (by contradiction) that this finite termination always occurs. We therefore assume that the loop is infinite and j tends to infinity.

Suppose first that $t_{\max} = \infty$ for all $j \ge 0$. Because of Lemma 3.4, we know that $\theta(x_k, t_j) = \|x_{k,j} - x_k\|$ is bounded above as a function of j, but yet $t_{j+1} = 2t_j$ and thus t_j tends to infinity. We may then apply (3.13) to deduce that

$$\|P_{\mathcal{T}(x_{k,j})}[-g_k]\| \le \frac{\theta(x_k, t_j)}{t_j},$$

and thus that

$$\lim_{j\to\infty} \|P_{\mathcal{T}(x_{k,j})}[-g_k]\| = 0. \tag{3.24}$$

But Theorem 12.1.4 of Conn et al. (2000) gives that, for all $j \ge 0$,

$$-\langle g_k, x_{k,j} - x_k\rangle = |\langle g_k, x_{k,j} - x_k\rangle| = \chi(x_k, \|x_{k,j} - x_k\|),$$

and therefore, using Lemma 3.1, that $|\langle g_k, x_{k,j} - x_k\rangle|$ is non-decreasing with j and that

$$|\langle g_k, x_{k,0} - x_k\rangle| = \chi(x_k, \|x_{k,0} - x_k\|) \ge \min[1, \|x_{k,0} - x_k\|]\,\chi_k > 0,$$

where the last inequality follows from the fact that x_k is not first-order critical. As a consequence,

$$|\langle g_k, x_{k,j} - x_k\rangle| \ge \min[1, \|x_{k,0} - x_k\|]\,\chi_k > 0$$

for all $j \ge 0$. Combining this observation with (3.24), we conclude that (2.6) must hold for all j sufficiently large, and the loop inside Step 1 must then be finite, which contradicts our assumption. Thus our initial supposition on $t_{\max}$ is impossible and $t_{\max}$ must be reset to a finite value. The continuity of the model m_k and of the projection operator $P_{\mathcal{F}}$ then imply, together with (2.7), the existence of an interval I of $\mathbb{R}^+$ of nonzero length such that, for all $t \in I$,

$$m_k(P_{\mathcal{F}}[x_k - tg_k]) \le f(x_k) + \kappa_{\text{ubs}}\langle g_k, P_{\mathcal{F}}[x_k - tg_k] - x_k\rangle$$

and

$$m_k(P_{\mathcal{F}}[x_k - tg_k]) \ge f(x_k) + \kappa_{\text{lbs}}\langle g_k, P_{\mathcal{F}}[x_k - tg_k] - x_k\rangle.$$

But this interval is independent of j and is always contained in $[t_{\min}, t_{\max}]$ by construction, while the length of this latter interval converges to zero when j tends to infinity. Hence there must exist a finite j such that both (2.4) and (2.5) hold, leading to the desired contradiction. □

We now derive two finer upper bounds on the length of the generalized Cauchy step, depending on two different criticality measures. These results are inspired by Lemma 2.1 of Cartis et al. (2007).

Lemma 3.6 *Suppose that AS1 and AS2 hold. Then we have that*

$$\|s_k^{GC}\| \le \frac{3}{\sigma_k}\max\left[\|B_k\|, (\sigma_k\chi_k)^{\frac{1}{2}}, \left(\sigma_k^2\chi_k\right)^{\frac{1}{3}}\right]. \tag{3.25}$$

and

$$\|s_k^{GC}\| \le \frac{3}{\sigma_k}\max\left[\|B_k\|, (\sigma_k\pi_k^{GC})^{\frac{1}{2}}\right]. \tag{3.26}$$

Proof. For brevity, we omit the index k. From (2.2), (3.12) and the Cauchy-Schwarz inequality,

$$\begin{aligned} m(x^{\text{GC}}) - f(x) &= \langle g, s^{\text{GC}}\rangle + \tfrac{1}{2}\langle s^{\text{GC}}, Bs^{\text{GC}}\rangle + \tfrac{1}{3}\sigma\|s^{\text{GC}}\|^3 \\ &\ge -\chi(x, \|s^{\text{GC}}\|) - \tfrac{1}{2}\|s^{\text{GC}}\|^2\|B\| + \tfrac{1}{3}\sigma\|s^{\text{GC}}\|^3 \\ &= \left[\tfrac{1}{9}\sigma\|s^{\text{GC}}\|^3 - \chi(x, \|s^{\text{GC}}\|)\right] + \left[\tfrac{2}{9}\sigma\|s^{\text{GC}}\|^3 - \tfrac{1}{2}\|s^{\text{GC}}\|^2\|B\|\right]. \end{aligned} \tag{3.27}$$

Thus since $m(x^{\text{GC}}) \le f(x)$, at least one of the bracketed expressions must be negative, i.e. either

$$\|s^{\text{GC}}\| \le \frac{9}{4}\frac{\|B\|}{\sigma} \tag{3.28}$$

or

$$\|s^{\text{GC}}\|^3 \le \frac{9}{\sigma}\chi(x, \|s^{\text{GC}}\|); \tag{3.29}$$

the latter is equivalent to

$$\|s^{\text{GC}}\| \le 3\left(\frac{\pi^{\text{GC}}}{\sigma}\right)^{\frac{1}{2}} \tag{3.30}$$

because of (3.5) when $\theta = \|s^{\text{GC}}\|$. In the case that $\|s^{\text{GC}}\| \ge 1$, (3.7) then gives that

$$\|s^{\text{GC}}\| \le 3\left(\frac{\chi}{\sigma}\right)^{\frac{1}{2}}. \tag{3.31}$$

Conversely, if $\|s^{\text{GC}}\| < 1$, we obtain from (3.8) and (3.29) that

$$\|s^{\text{GC}}\| \le 3\left(\frac{\chi}{\sigma}\right)^{\frac{1}{3}}. \tag{3.32}$$

Gathering (3.28), (3.31) and (3.32), we immediately obtain (3.25). Combining (3.28) and (3.30) gives (3.26). □

Similar results may then be derived for the length of the full step, as we now show.

Lemma 3.7 *Suppose that AS1 and AS2 hold, and that*

$$\|s_k\| \le \frac{3}{\sigma_k}\max\left[\|B_k\|, (\sigma_k\chi_k)^{\frac{1}{2}}, \left(\sigma_k^2\chi_k\right)^{\frac{1}{3}}\right] \tag{3.33}$$

and

$$\|s_k\| \le \frac{3}{\sigma_k}\max\left[\|B_k\|, \sqrt{\sigma_k\pi_k^{GC}}\right]. \tag{3.34}$$

Proof. We start by proving (3.33) and

$$\|s_k\| \le \frac{3}{\sigma_k} \max\left[\|B_k\|, \sqrt{\sigma_k \pi_k^+}\right] \tag{3.35}$$

in a manner identical to that used for (3.25) and (3.26) with s_k replacing $s_k^{\rm GC}$: we now use the inequality $\langle g_k, s_k\rangle \ge -\chi(x_k, \|s_k\|)$ (itself resulting from (3.1)) in (3.27) instead of (3.12), and also (3.9),(3.10) instead of (3.7),(3.8) to derive the analogues of (3.31) and (3.32). Now, if $\|s_k\| \le \|s_k^{\rm GC}\|$, then (3.26) gives that (3.34) holds. Otherwise, i.e. if $\|s_k\| > \|s_k^{\rm GC}\|$, then the non-increasing nature of $\pi(x_k, \theta)$ gives that $\pi_k^+ \le \pi_k^{\rm GC}$. Substituting this inequality in (3.35) also gives (3.34). □

Using the above results, we may then derive the equivalent of the well-known Cauchy decrease condition in our contrained case. Again the exact expression of this condition depends on the criticality measure considered.

Lemma 3.8 *If $\|s_k^{GC}\| \ge 1$, then, for $\kappa_{\rm GC} \stackrel{\rm def}{=} \min[\frac{1}{2}, \frac{2}{3}\kappa_{\rm ubs}(1-\kappa_{\rm lbs})] \in (0,1)$,*

$$m_k(x_k) - m_k(x_k^{GC}) \ge \kappa_{\rm GC}\chi_k. \tag{3.36}$$

If $\|s_k^{GC}\| \le 1$, then

$$m_k(x_k) - m_k(x_k^{GC}) \ge \kappa_{\rm GC}\pi_k^{GC} \min\left[\frac{\pi_k^{GC}}{1+\|B_k\|}, \sqrt{\frac{\pi_k^{GC}}{\sigma_k}}\right] \tag{3.37}$$

if (2.5) holds, while

$$m_k(x_k) - m_k(x_k^{GC}) \ge \kappa_{\rm GC}\chi_k \min\left[\frac{\chi_k}{1+\|B_k\|}, \sqrt{\frac{\pi_k^{GC}}{\sigma_k}}, 1\right] \tag{3.38}$$

if (2.5) fails. In all cases,

$$m_k(x_k) - m_k(x_k^{GC}) \ge \kappa_{\rm GC}\chi_k \min\left[\frac{\chi_k}{1+\|B_k\|}, \sqrt{\frac{\chi_k}{\sigma_k}}, 1\right]. \tag{3.39}$$

Proof. Again, we omit the index k for brevity. First note that, because of (2.4) and (3.12),

$$f(x) - m(x^{\rm GC}) \ge \kappa_{\rm ubs}|\langle g, s^{\rm GC}\rangle| = \kappa_{\rm ubs}\chi(x, \|s^{\rm GC}\|) = \kappa_{\rm ubs}\pi(x, \|s^{\rm GC}\|)\|s^{\rm GC}\|. \tag{3.40}$$

Assume first that $\|s^{\rm GC}\| \ge 1$. Then, using (3.7), we see that

$$f(x) - m(x^{\rm GC}) \ge \kappa_{\rm ubs}\chi, \tag{3.41}$$

which gives (3.36) since $\kappa_{\rm ubs} > \kappa_{\rm GC}$. Assume now, for the remainder of the proof, that $\|s^{\rm GC}\| \le 1$, which implies, by (3.8), that

$$f(x) - m(x^{\rm GC}) \ge \kappa_{\rm ubs}\chi\|s^{\rm GC}\|, \tag{3.42}$$

and first consider the case where (2.5) holds. Then, from (2.2) and (2.5), the Cauchy-Schwarz inequality, (3.12) and (3.5), we obtain that

$$\|B\| + \tfrac{2}{3}\sigma\|s^{\rm GC}\| \ge \frac{2(1-\kappa_{\rm lbs})}{\|s^{\rm GC}\|^2}|\langle g, s^{\rm GC}\rangle| = \frac{2(1-\kappa_{\rm lbs})}{\|s^{\rm GC}\|^2}\chi(x, \|s^{\rm GC}\|) = \frac{2(1-\kappa_{\rm lbs})}{\|s^{\rm GC}\|}\pi^{\rm GC}$$

and hence that

$$\|s^{\rm GC}\| \ge \frac{2(1-\kappa_{\rm lbs})\pi^{\rm GC}}{\|B\| + \frac{2}{3}\sigma\|s^{\rm GC}\|}.$$

Recalling (3.26), we thus deduce that

$$\|s^{\rm GC}\| \ge \frac{2(1-\kappa_{\rm lbs})\pi^{\rm GC}}{\|B\| + 2\max\left[\|B\|, \sqrt{\sigma\pi^{\rm GC}}\right]}.$$

Combining this inequality with (3.40), we obtain that

$$f(x) - m(x^{\rm GC}) \geq \tfrac{2}{3}\kappa_{\rm ubs}(1-\kappa_{\rm lbs})\pi^{\rm GC}\min\left[\frac{\pi^{\rm GC}}{1+\|B\|}, \sqrt{\frac{\pi^{\rm GC}}{\sigma}}\right],$$

which implies (3.37).

If (2.5) does not hold (and $\|s_k^{\rm GC}\| \leq 1$), then (2.6) must hold. Thus, (3.11) and (2.7) imply that

$$\chi \leq (1+2\kappa_{\rm epp})\chi(x, \|s^{\rm GC}\|) \leq 2\chi(x, \|s^{\rm GC}\|).$$

Substituting this inequality in (3.40) then gives that

$$f(x) - m(x^{\rm GC}) \geq \tfrac{1}{2}\kappa_{\rm ubs}\chi. \tag{3.43}$$

This in turn gives (3.36). The inequality (3.38) results from (3.37) and (3.8), in the case when (2.5) holds, and (3.43) when (2.5) does not hold. Finally, (3.39) follows from combining (3.37) and (3.36) and using (3.8) in the former. □

We next show that when the iterate x_k is sufficiently non-critical, then iteration k must be very successful and the regularisation parameter does not increase.

Lemma 3.9 *Suppose AS1–AS3 hold, that $\chi_k > 0$ and that*

$$\min\left[\sigma_k, (\sigma_k\chi_k)^{\frac{1}{2}}, (\sigma_k^2\chi_k)^{\frac{1}{3}}\right] \geq \frac{9(\kappa_{\rm H}+\kappa_{\rm B})}{2(1-\eta_2)\kappa_{\rm GC}} \stackrel{\rm def}{=} \kappa_{\rm suc} > 1. \tag{3.44}$$

Then iteration k is very successful and

$$\sigma_{k+1} \leq \sigma_k. \tag{3.45}$$

Proof. First note that the last inequality in (3.44) follows from the facts that $\kappa_{\rm H} \geq 1$, $\kappa_{\rm B} \geq 1$ and $\kappa_{\rm GC} \in (0,1)$. Again, we omit the index k for brevity. The mean-value theorem gives that

$$f(x^+) - m(x^+) = \tfrac{1}{2}\langle s, [H(\xi) - B]s\rangle - \tfrac{1}{3}\sigma\|s\|^3$$

for some $\xi \in [x, x^+]$. Hence, using (3.2),

$$f(x^+) - m(x^+) \leq \tfrac{1}{2}(\kappa_{\rm H}+\kappa_{\rm B})\|s\|^2. \tag{3.46}$$

We also note that (3.44) and AS4 imply that $(\sigma\chi)^{\frac{1}{2}} \geq \|B\|$ and hence, from (3.33), that

$$\|s\| \leq \frac{3}{\sigma}\max\left[(\sigma\chi)^{\frac{1}{2}}, (\sigma^2\chi)^{\frac{1}{3}}\right] = 3\max\left[\left(\frac{\chi}{\sigma}\right)^{\frac{1}{2}}, \left(\frac{\chi}{\sigma}\right)^{\frac{1}{3}}\right].$$

Substituting this last bound in (3.46) then gives that

$$f(x^+) - m(x^+) \leq \frac{9(\kappa_{\rm H}+\kappa_{\rm B})}{2}\max\left[\frac{\chi}{\sigma}, \left(\frac{\chi}{\sigma}\right)^{\frac{2}{3}}\right]. \tag{3.47}$$

Assume now that $\|s^{\rm GC}\| \leq 1$ and that (2.6) holds but not (2.5), or that $\|s^{\rm GC}\| > 1$. Then (2.9) and (3.36) also imply that

$$f(x) - m(x^+) \geq f(x) - m(x^{\rm GC}) \geq \kappa_{\rm GC}\chi.$$

Thus, using this bound and (3.47),

$$\begin{aligned}
1-\rho &= \frac{f(x^+) - m(x^+)}{f(x) - m(x^+)} \\
&\leq \frac{9(\kappa_{\rm H}+\kappa_{\rm B})}{2\kappa_{\rm GC}\chi}\max\left[\frac{\chi}{\sigma}, \left(\frac{\chi}{\sigma}\right)^{\frac{2}{3}}\right] \\
&= \frac{9(\kappa_{\rm H}+\kappa_{\rm B})}{2\kappa_{\rm GC}}\max\left[\frac{1}{\sigma}, \frac{1}{(\sigma^2\chi)^{\frac{1}{3}}}\right] \\
&\leq 1-\eta_2
\end{aligned} \tag{3.48}$$

where the last inequality results from (3.44). Assume alternatively that $\|s^{\rm GC}\| \leq 1$ and (2.5) holds. We then deduce from (3.8), (3.44) and (3.2) that

$$\sqrt{\sigma\pi^{\rm GC}} \geq \sqrt{\sigma\chi} \geq \|B\|.$$

Then (3.34) yields that

$$\|s\| \leq 3\sqrt{\frac{\pi^{\rm GC}}{\sigma}},$$

which can be substituted in (3.46) to give that

$$f(x^+) - m(x^+) \leq \frac{9}{2}(\kappa_{\rm H} + \kappa_{\rm B})\frac{\pi^{\rm GC}}{\sigma}. \tag{3.49}$$

On the other hand, (2.9), (3.37) and (3.44) also imply that

$$f(x) - m(x^+) \geq f(x) - m(x^{\rm GC}) \geq \kappa_{\rm GC}\pi^{\rm GC}\sqrt{\frac{\pi^{\rm GC}}{\sigma}}.$$

Thus, using this last bound, (2.8), (3.49), (3.8) and (3.44), we obtain that

$$1 - \rho = \frac{f(x^+) - m(x^+)}{f(x) - m(x^+)} \leq \frac{9(\kappa_{\rm H} + \kappa_{\rm B})}{2\kappa_{\rm GC}\sqrt{\sigma\pi^{\rm GC}}} \leq \frac{9(\kappa_{\rm H} + \kappa_{\rm B})}{2\kappa_{\rm GC}\sqrt{\sigma\chi}} \leq 1 - \eta_2. \tag{3.50}$$

We then conclude from (3.48) and (3.50) that $\rho \geq \eta_2$ whenever (3.44) holds, which means that the iteration is very successful and (3.45) follows. □

Our next result shows that the regularisation parameter must remain bounded unless a critical point is approached. Note that this result does not depend on the objective's Hessian being Lipschitz continuous.

Lemma 3.10 *Suppose that AS1–AS3 hold, and that there is a constant $\epsilon \in (0,1]$ such that*

$$\chi_k \geq \epsilon \tag{3.51}$$

for all $k \geq 0$. Then, for all $k \geq 0$,

$$\sigma_k \leq \max\left[\sigma_0, \frac{\gamma_2\kappa_{\rm suc}^2}{\epsilon}\right] \stackrel{\rm def}{=} \kappa_\sigma. \tag{3.52}$$

Proof. Assume that

$$\sigma_k \geq \frac{\kappa_{\rm suc}^2}{\epsilon}. \tag{3.53}$$

Then $\sigma_k \geq \kappa_{\rm suc}$ because $\kappa_{\rm suc} > 1$ and $\epsilon < 1$. Moreover, one verifies easily, using (3.51), that

$$(\sigma_k\chi_k)^{\frac{1}{2}} \geq (\sigma_k\epsilon)^{\frac{1}{2}} = \left(\kappa_{\rm suc}^2\right)^{\frac{1}{2}} = \kappa_{\rm suc}$$

and that

$$\left(\sigma_k^2\chi_k\right)^{\frac{1}{3}} \geq \left(\frac{\kappa_{\rm suc}^4}{\epsilon}\right)^{\frac{1}{3}} \geq \left(\kappa_{\rm suc}^3\right)^{\frac{1}{3}} = \kappa_{\rm suc}.$$

Hence we deduce that, for each k, (3.53) implies that (3.44) holds. Hence, (3.53) ensures (3.45) because of Lemma 3.9. Thus, when $\sigma_0 \leq \gamma_2\kappa_{\rm suc}^2/\epsilon$, one also obtains that $\sigma_k \leq \gamma_2\kappa_{\rm suc}^2/\epsilon$ for all k, where we have introduced the factor γ_2 for the case where σ_k is less that $\kappa_{\rm suc}^2/\epsilon$ and iteration k is not very successful. Thus (3.52) holds. If, on the other hand, $\sigma_0 > \gamma_2\kappa_{\rm suc}^2/\epsilon$, the above reasoning shows that σ_k cannot increase, and (3.52) also holds. □

We are now ready to prove our first-order convergence result. We first state it for the case where there are only finitely many successful iterations.

Lemma 3.11 *Suppose that AS1–AS3 hold and that there are only finitely many successful iterations. Then $x_k = x_*$ for all sufficiently large k and x_* is first-order critical.*

Proof. See Lemma 2.5 in Cartis et al. (2007), with χ_k replacing $\|g_k\|$. □

We conclude this section by showing the desired convergence when the number of successful iterations is infinite. As for trust-region methods, this is accomplished by first showing first-order criticality along a subsequence of iterations.

Theorem 3.12 *Suppose that AS1–AS4 hold. Then we have that*

$$\liminf_{k\to\infty} \chi_k = 0. \tag{3.54}$$

Hence, at least one limit point of the sequence $\{x_k\}$ (if any) is first-order critical.

Proof. The conclusion holds when there are finitely many successful iterations because of Lemma 3.11. Suppose therefore that there are infinitely many successful iterations. Suppose furthermore that (3.51) holds for all k. The mechanism of the algorithm then implies that, if iteration k is successful,

$$f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(x_k^+)] \geq \eta_1 \kappa_{\rm GC} \chi_k \min\left[\frac{\chi_k}{1+\|B_k\|}, \sqrt{\frac{\chi_k}{\sigma_k}}, 1\right],$$

where we have used (2.9) and (3.39) to obtain the last inequality. The bounds (3.2), (3.51) and (3.52) then yield that

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \kappa_{\rm GC} \epsilon \min\left[\frac{\epsilon}{1+\kappa_{\rm B}}, \sqrt{\frac{\epsilon}{\kappa_\sigma}}, 1\right] \stackrel{\rm def}{=} \kappa_\epsilon > 0. \tag{3.55}$$

Summing over all successful iterations, we deduce that

$$f(x_0) - f(x_{k+1}) = \sum_{j=0, j\in\mathcal{S}}^{k} [f(x_j) - f(x_{j+1})] \geq i_k \kappa_\epsilon,$$

where i_k is the number of successful iterations up to iteration k. Since i_k tends to infinity by assumption, we obtain that the sequence $\{f(x_k)\}$ tends to minus infinity, which is impossible because f is bounded below on $\mathcal{F}$ and $x_k \in \mathcal{F}$ for all k. Hence (3.51) cannot hold and (3.54) follows. □

We finally prove that the conclusion of the last theorem is not restricted to a subsequence, but holds for the complete sequence of iterates.

Theorem 3.13 *Suppose that AS1–AS4 hold. Then we have that*

$$\lim_{k\to\infty} \chi_k = 0, \tag{3.56}$$

and all limit points of the sequence $\{x_k\}$ (if any) are first-order critical.

Proof. If $\mathcal{S}$ is finite, the conclusion directly follows from Lemma 3.11. Suppose therefore that there are infinitely many successful iterations and that there exists a subsequence $\{t_i\} \subseteq \mathcal{S}$ such that

$$\chi_{t_i} \geq 2\epsilon \tag{3.57}$$

for some $\epsilon > 0$. From (3.54), we deduce the existence of another subsequence $\{\ell_i\} \subseteq \mathcal{S}$ such that, for all i, ℓ_i is the index of the first successful iteration after iteration t_i such that

$$\chi_k \geq \epsilon \;\text{ for }\; t_i \leq k < \ell_i \;\text{ and }\; \chi_{\ell_i} \leq \epsilon. \tag{3.58}$$

We then define

$$\mathcal{K} = \{k \in \mathcal{S} \mid t_i \leq k < \ell_i\}. \tag{3.59}$$

Thus, for each $k \in \mathcal{K} \subseteq \mathcal{S}$, we obtain from (3.39) and (3.58) that

$$f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(x_k^+)] \geq \eta_1 \kappa_{\rm GC} \epsilon \min\left[\frac{\epsilon}{1+\|B_k\|}, \sqrt{\frac{\chi_k}{\sigma_k}}, 1\right]. \tag{3.60}$$

Because $\{f(x_k)\}$ is monotonically decreasing and bounded below, it must be convergent and we thus deduce from (3.60) that

$$\lim_{k\to\infty, k\in\mathcal{K}} \frac{\chi_k}{\sigma_k} = 0, \tag{3.61}$$

which in turn implies, in view of (3.58), that

$$\lim_{k\to\infty, k\in\mathcal{K}} \sigma_k = +\infty. \tag{3.62}$$

As a consequence of this limit, (3.25), (3.2) and (3.58), we see that, for $k \in \mathcal{K}$,

$$\|s_k^{\rm GC}\| \leq 3 \max\left[\frac{\kappa_{\rm B}}{\sigma_k}, \left(\frac{\chi_k}{\sigma_k}\right)^{\frac{1}{2}}, \left(\frac{\chi_k}{\sigma_k}\right)^{\frac{2}{3}}\right],$$

and thus $\|s_k^{\rm GC}\|$ converges to zero along $\mathcal{K}$. We therefore obtain that

$$\|s_k^{\rm GC}\| < 1 \text{ for all } k \in \mathcal{K} \text{ sufficiently large}, \tag{3.63}$$

which implies that (3.38) is applicable for these k, yielding, in view of (3.2) and (3.58), that, for $k \in \mathcal{K}$ sufficiently large,

$$f(x_k) - f(x_{k+1}) \geq \eta_1[f(x_k) - m_k(x_k^+)] \geq \eta_1 \kappa_{\rm GC} \epsilon \min\left[\frac{\epsilon}{1+\kappa_{\rm B}}, \sqrt{\frac{\pi_k^{\rm GC}}{\sigma_k}}, 1\right],$$

where we have used (3.8), (3.61) and (3.63) to deduce the last inequality. But the convergence of the sequence $\{f(x_k)\}$ implies that the left-hand side of this inequality converges to zero, and hence that the minimum in the last right-hand side must be attained by its middle term for $k \in \mathcal{K}$ sufficiently large. We therefore deduce that, for these k,

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \kappa_{\rm GC} \epsilon \sqrt{\frac{\pi_k^{\rm GC}}{\sigma_k}}. \tag{3.64}$$

We also obtain from (3.8) that $\pi_k^{\rm GC} \geq \chi_k \geq \epsilon$. As a consequence, (3.34), (3.2) and (3.62) ensure that

$$\|s_k\| \leq 3\sqrt{\frac{\pi_k^{\rm GC}}{\sigma_k}} \leq \frac{3}{\eta_1 \kappa_{\rm GC} \epsilon}[f(x_k) - f(x_{k+1})]$$

for $k \in \mathcal{K}$ sufficiently large. This last bound can then be used to see that

$$\|x_{\ell_i} - x_{t_i}\| \leq \frac{3}{\eta_1 \kappa_{\rm GC} \epsilon} \sum_{k=t_i, k\in\mathcal{K}}^{\ell_i - 1} [f(x_k) - f(x_{k+1})] \leq \frac{3}{\eta_1 \kappa_{\rm GC} \epsilon}[f(x_{t_i}) - f(x_{\ell_i})].$$

Since $\{f(x_k)\}$ is convergent, the right-hand side of this inequality tends to zero as i tends to infinity. Hence $\|x_{\ell_i} - x_{t_i}\|$ converges to zero with i, and, by continuity, so does $\|\chi_{\ell_i} - \chi_{t_i}\|$. But this is impossible in view of (3.57) and (3.58). Hence no subsequence can exist such that (3.57) holds and the proof is complete. □

4 Worst-Case Function-Evaluation Complexity

This section is devoted to worst-case function-evaluation complexity bounds, that is bounds on the number of objective function or gradient evaluations needed to achieve first-order convergence to prescribed accuracy. Despite the obvious observation that such an analysis does not cover the total computational cost of solving a problem, this type of complexity result is of special interest for nonlinear optimization because there are many examples where the cost these evaluations completely dwarfs that of the other computations inside of the algorithm itself.

4.1 Function-Evaluation Complexity for Algorithm ACURC

We first consider the function- (and gradient-) evaluation complexity of a variant (ACUR$_\epsilon$) of the ACURC algorithm itself, only differing by the introduction of an approximate termination rule. More specifically, we replace the criticality check in Step 1 of ACURC by the test $\chi_k \leq \epsilon$ (where ϵ is user-supplied threshold) and terminate if this inequality holds. The results presented for this algorithm are inspired by complexity results for trust-region algorithms (see Gratton, Sartenaer and Toint, 2008*b*, Gratton et al., 2008*a*) and for the adaptive cubic overestimation algorithm (see Cartis et al., 2007).

Theorem 4.1 *Suppose that AS1-AS3 hold, that, for all $k \in \mathcal{S}$ and some $\gamma_3 \in (0,1)$,*

$$\sigma_{k+1} \geq \gamma_3 \sigma_k \quad \textit{whenever} \quad \rho_k \geq \eta_2, \tag{4.1}$$

and that the approximate criticality threshold ϵ is small enough to ensure

$$\epsilon \leq \min\left[1, \frac{\gamma_2 \kappa_{\rm suc}^2}{\sigma_0}\right]. \tag{4.2}$$

Then there exists a constant $\kappa_{\rm df} \in (0,1)$ such that, for every $k \geq 0$, $k \in \mathcal{S}$,

$$f(x_k) - f(x_{k+1}) \geq \kappa_{\rm df} \epsilon^2 \tag{4.3}$$

before Algorithm ACURC$_\epsilon$ terminates, that is generates an iterate x_k such that $\chi_k \leq \epsilon$. As a consequence, this algorithm needs at most

$$\left\lceil \kappa_{\mathcal{S}} \epsilon^{-2} \right\rceil$$

successful iterations and evaluations of $\nabla_x f$, and at most

$$\left\lceil \kappa_* \epsilon^{-2} \right\rceil$$

iterations and objective function evaluations to terminate, where

$$\kappa_{\mathcal{S}} \stackrel{\rm def}{=} \left\lceil \frac{f(x_0) - f_{\rm low}}{\eta_1 \kappa_{\rm df}} \right\rceil \quad \textit{and} \quad \kappa_* \stackrel{\rm def}{=} \left[1 - \frac{\log(\gamma_3)}{\gamma_1}\right] \kappa_{\mathcal{S}} + \frac{1}{\log(\gamma_1)} \max\left[1, \frac{\gamma_2 \kappa_{\rm suc}^2}{\sigma_0}\right].$$

Proof. We first note that, as long as Algorithm ACURC$_\epsilon$ has not terminated, $\chi_k > \epsilon$. We may then use the same reasoning as in the proof of Theorem 3.12 and use (3.52) and (3.55) to deduce that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta_1 \kappa_{\rm GC} \epsilon \min\left[\frac{\epsilon}{1+\kappa_{\rm B}}, \sqrt{\frac{\epsilon}{\max\left[\sigma_0, \gamma_2 \kappa_{\rm suc}^2/\epsilon\right]}}, 1\right] \\ &\geq \eta_1 \kappa_{\rm GC} \min\left[\frac{1}{1+\kappa_{\rm H}}, \frac{1}{\kappa_{\rm suc}\sqrt{\gamma_2}}\right] \epsilon^2 \end{aligned}$$

where we have used (4.2) to derive the last inequality. This gives (4.3) with

$$\kappa_{\rm df} \stackrel{\rm def}{=} \eta_1 \kappa_{\rm GC} \min\left[\frac{1}{1+\kappa_{\rm H}}, \frac{1}{\kappa_{\rm suc}\sqrt{\gamma_2}}\right].$$

The proof is then completed by using Theorem 6.4 in Cartis et al. (2007). □

Because Algorithm ACURC does not exploit more than first-order information (via the Cauchy point definition), this bound is, as expected, similar in nature to that obtained by Nesterov (2004), page 29, for the steepest descent method.

4.2 An $O(\epsilon^{-\frac{3}{2}})$ Function-Evaluation Complexity Bound

We now discuss a close variant (ACURC-S) of the ACURC algorithm for which an interesting worst-case function- (and derivatives-) evaluation complexity result can be shown. Algorithm ACURC-S uses the user-supplied first-order accuracy threshold $\epsilon > 0$. It differs from Algorithm ACURC in that stronger conditions are imposed on the step.

We first prove the following useful technical lemma.

Lemma 4.2 *Suppose that*

$$\langle \nabla_x m_k(x_k^+), s_k \rangle \le 0. \tag{4.4}$$

and that

$$\langle \nabla_x m_k(x_k), s_k \rangle \le 0 \quad \textit{or} \quad \langle s_k, B_k s_k \rangle \ge 0. \tag{4.5}$$

Then

$$f(x_k) - m_k(x_k^+) \ge \tfrac{1}{6}\sigma_k \|s_k\|^3. \tag{4.6}$$

Proof. (Dropping the index k again.) Condition (4.4) is equivalent to

$$\langle g, s \rangle + \langle s, Bs \rangle + \sigma \|s\|^3 \le 0. \tag{4.7}$$

If $\langle s, Bs \rangle \ge 0$, we substitute $\langle g, s \rangle$ from this inequality in (2.2) and deduce that

$$m(x^+) - f(x) = \langle g, s \rangle + \tfrac{1}{2}\langle s, Bs \rangle + \tfrac{1}{3}\sigma\|s\|^3 \le -\tfrac{1}{2}\langle s, Bs \rangle - \tfrac{2}{3}\sigma\|s\|^3,$$

which then implies (4.6). If, on the other hand, $\langle s, Bs \rangle < 0$, then we substitute the inequality on $\langle s, Bs \rangle$ resulting from (4.7) into (2.2) and obtain that

$$m(x^+) - f(x) = \langle g, s \rangle + \tfrac{1}{2}\langle s, Bs \rangle + \tfrac{1}{3}\sigma\|s\|^3 \le \tfrac{1}{2}\langle g, s \rangle - \tfrac{1}{6}\sigma\|s\|^3,$$

from which (4.6) again follows because of (4.5). □

Thus, as long as the step is along a descent or non-negative curvature direction, the model decrease is bounded below by a fraction of the norm of the step cubed. This result may be extended as follows.

Lemma 4.3 *Suppose that there exist steps $s_{k,\circ}$ and $s_{k,\bullet}$ and points $x_{k,\circ} = x_k + s_{k,\circ}$ and $x_{k,\bullet} = x_k + s_{k,\bullet}$ such that, for some $\kappa \in (0,1]$,*

$$m_k(x_{k,\circ}) \le m_k(x_k) - \kappa\sigma_k\|s_{k,\circ}\|^3, \tag{4.8}$$

$$m_k(x_{k,\bullet}) \le m_k(x_{k,\circ}), \tag{4.9}$$

$$\langle \nabla_x m_k(x_{k,\bullet}), x_{k,\bullet} - x_{k,\circ} \rangle \le 0, \tag{4.10}$$

and

$$\langle \nabla_x m_k(x_{k,\circ}), x_{k,\bullet} - x_{k,\circ} \rangle \le 0. \tag{4.11}$$

Then

$$m_k(x_k) - m_k(x_{k,\bullet}) \ge \kappa_{\rm lm}\kappa\sigma_k\|s_{k,\bullet}\|^3. \tag{4.12}$$

for some constant $\kappa_{\rm lm} \in (0,1)$ independent of k and κ.

Proof. (Dropping the index k again.) Suppose first that, for some $\alpha \in (0,1)$,

$$\|s_\circ\| \ge \alpha\|s_\bullet\|. \tag{4.13}$$

Then (4.8) and (4.9) give that

$$m(x) - m(x_\bullet) = m(x) - m(x_\circ) + m(x_\circ) - m(x_\bullet) \ge \kappa\sigma\|s_\circ\|^3 \ge \kappa\sigma_k\alpha^3\|s_\bullet\|^3. \tag{4.14}$$

Assume now that (4.13) fails, that is

$$\|s_\circ\| < \alpha\|s_\bullet\|. \tag{4.15}$$

We have that

$$f(x) + \langle g, s_\circ \rangle + \tfrac{1}{2}\langle s_\circ, Bs_\circ \rangle = m(x_\circ) - \tfrac{1}{3}\sigma\|s_\circ\|^3. \tag{4.16}$$

Using this identity, we now see that

$$\begin{aligned} m(x_\bullet) &= f(x) + \langle g, s_\circ \rangle + \tfrac{1}{2}\langle s_\circ, Bs_\circ \rangle + \langle g + Bs_\circ, s_\bullet - s_\circ \rangle + \tfrac{1}{2}\langle s_\bullet - s_\circ, B(s_\bullet - s_\circ) \rangle + \tfrac{1}{3}\sigma\|s_\bullet\|^3 \\ &= m(x_\circ) + \langle g + Bs_\circ, s_\bullet - s_\circ \rangle + \tfrac{1}{2}\langle s_\bullet - s_\circ, B(s_\bullet - s_\circ) \rangle + \tfrac{1}{3}\sigma\|s_\bullet\|^3 - \tfrac{1}{3}\sigma\|s_\circ\|^3 \end{aligned} \tag{4.17}$$

Moreover, (4.10) yields that

$$0 \geq \langle g+Bs_\bullet, s_\bullet - s_\circ\rangle + \sigma\|s_\bullet\|\langle s_\bullet, s_\bullet - s_\circ\rangle = \langle g+Bs_\circ, s_\bullet - s_\circ\rangle + \langle s_\bullet - s_\circ, B(s_\bullet - s_\circ)\rangle + \sigma\|s_\bullet\|\langle s_\bullet, s_\bullet - s_\circ\rangle,$$

and thus (4.17) becomes

$$m(x_\bullet) \leq m(x_\circ) + \tfrac{1}{2}\langle g + Bs_\circ, s_\bullet - s_\circ\rangle - \tfrac{1}{2}\sigma\|s_\bullet\|\langle s_\bullet, s_\bullet - s_\circ\rangle + \tfrac{1}{3}\sigma\|s_\bullet\|^3 - \tfrac{1}{3}\sigma\|s_\circ\|^3. \tag{4.18}$$

But we may also use (4.11) and deduce that

$$0 \geq \langle g + Bs_\circ, s_\bullet - s_\circ\rangle + \sigma\|s_\circ\|\langle s_\circ, s_\bullet - s_\circ\rangle,$$

which, together with (4.18), gives that

$$\begin{aligned} m(x_\circ) - m(x_\bullet) &\geq \tfrac{1}{2}\sigma\|s_\circ\|\langle s_\circ, s_\bullet - s_\circ\rangle + \tfrac{1}{2}\sigma\|s_\bullet\|\langle s_\bullet, s_\bullet - s_\circ\rangle - \tfrac{1}{3}\sigma\|s_\bullet\|^3 + \tfrac{1}{3}\sigma\|s_\circ\|^3 \\ &\geq \sigma\big(-\tfrac{1}{2}\|s_\circ\|^2\|s_\bullet\| - \tfrac{1}{6}\|s_\circ\|^3 + \tfrac{1}{6}\|s_\bullet\|^3 - \tfrac{1}{2}\|s_\bullet\|^2\|s_\circ\|\big), \end{aligned} \tag{4.19}$$

where we have used the Cauchy-Schwartz inequality. Taking now (4.8) and (4.15) into account and using the fact that $\kappa \leq 1$, we obtain that

$$m(x) - m(x_\bullet) \geq m(x_\circ) - m(x_\bullet) > \kappa\sigma\big(-\tfrac{1}{2}\alpha^2 - \tfrac{1}{6}\alpha^3 + \tfrac{1}{6} - \tfrac{1}{2}\alpha\big)\|s_\bullet\|^3. \tag{4.20}$$

We now select the value of α for which the lower bounds (4.14) and (4.20) are equal, namely $\alpha_* \approx 0.2418$, the only real positive root of $7\alpha^3 + 3\alpha^2 + 3\alpha = 1$. The desired result now follows from (4.14) and (4.20) with $\kappa_{\rm lm} \stackrel{\rm def}{=} \alpha_*^3 \approx 0.0141$. □

As it turns out, obtaining a lower bound of the type (4.6) or (4.12) is crucial for deriving the desired complexity result, as will become clear below. We thus need to ensure that our step computation ensures this property, which entails imposing further restrictions on the step. One first additional requirement is the following.

AS5: For all k, the step s_k solves the subproblem

$$\min_{s\in\mathbb{R}^n,\, x_k+s\in\mathcal{F}} m_k(x_k + s) \tag{4.21}$$

accurately enough to ensure that

$$\chi_k^{\rm m}(x_k^+, 1) \leq \min(\kappa_{\rm stop}, \|s_k\|)\, \chi_k \tag{4.22}$$

where $\kappa_{\rm stop} \in [0, 1)$ is a constant and where, for $\theta \geq 0$,

$$\chi_k^{\rm m}(x, \theta) \stackrel{\rm def}{=} \left| \min_{x+d\in\mathcal{F}, \|d\|\leq\theta} \langle \nabla_x m_k(x), d\rangle \right|. \tag{4.23}$$

The inequality (4.22) is an adequate stopping condition for the subproblem solution since $\chi_k^{\rm m}(x_k^*, 1)$ must be identically zero if x_k^* is a local minimizer of (4.21). It is the constrained analogue of the "s-stopping rule" of Cartis et al. (2007).

AS5 is however not sufficient for obtaining the desired result. As in Cartis et al. (2007) where (4.4) is imposed, one also needs to verify that a cheap model improvement cannot be obtained from x_k^+ for this point to be an acceptable trial point. However, at variance with the unconstrained case, there is no longer any guarantee that the step provides a descent direction (i.e. the first part of (4.5) holds). We therefore distinguish two possibilities. Assume first that (4.5) holds for the computed x_k^+. Then it is sufficient to require that (4.4) also holds. This condition expresses the reasonable requirement that the stepsize along s_k does not exceed that corresponding to the minimum of the model $m_k(x_k + \tau s_k)$ for $\tau > 0$. It is for instance satisfied if

$$\operatorname*{argmin}_{\tau\geq 0,\, x_k+\tau s_k\in\mathcal{F}} m_k(x_k + \tau s_k) = 1.$$

Note that (4.4) also holds at a local minimizer. The situation is more complicated when (4.5) fails, that is when the step is ascent (at x_k) rather than descent and of negative curvature. Our requirement on the trial point is then essentially that it can be computed by a uniformly bounded sequence of (possibly incomplete) line minimizations starting from x_k. More formally, we assume that there exist an integer $\bar{\ell} > 0$ and, for each k such that (4.5) fails, feasible points $\{x_{k,i}\}_{i=0}^{\ell_k}$ with $0 < \ell_k \leq \bar{\ell}$, $x_{k,0} = x_k$ and $x_{k,\ell_k} = x_k^+$, such that, for $i = 1, \ldots, \ell_k$,

$$m_k(x_{k,i}) \leq m_k(x_{k,i-1}), \quad \langle \nabla_x m_k(x_{k,i-1}), x_{k,i} - x_{k,i-1} \rangle \leq 0 \quad \text{and} \quad \langle \nabla_x m_k(x_{k,i}), x_{k,i} - x_{k,i-1} \rangle \leq 0. \tag{4.24}$$

Observe that these inequalities hold in particular if x_k^+ is the first minimizer of the model along the piecewise linear path

$$\mathcal{P}_k \stackrel{\text{def}}{=} \bigcup_{i=1}^{\ell_k} [x_{k,i-1}, x_{k,i}].$$

Observe also that (4.24) subsumes the condition discussed in the case where (4.5) holds, because one may then choose $\ell_k = 1$ and (4.24) then implies both (4.5) and (4.4). We therefore summarize these requirements in the form of

AS6: For all k, the step s_k is such that (4.24) holds for some $\{x_{k,i}\}_{i=0}^{\ell_k} \subset \mathcal{F}$ with $0 < \ell_k \leq \bar{\ell}$, $x_{k,0} = x_k$ and $x_{k,\ell_k} = x_k^+$.

Observe that we have not used global constrained optimization anywhere in the requirements imposed on the step s_k.

In practice, verifying AS6 need not be too burdensome. Firstly, the computation of x_k^+ may be by a sequence of line minimizations, and AS6 then trivially holds provided the number of such minimizations remains uniformly bounded. If the trial step has been determined by another technique, one might proceed as follows. If we set x_b to be the global minimum of the model in the hyperplane orthogonal to the gradient, that is

$$x_{k,b} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\langle g_k, s\rangle = 0} m_k(x_k + s), \tag{4.25}$$

then we may also define $x_{k,a}$ as the intersection of the segment $[x_k, x_{k,b}]$ with the boundary of $\mathcal{F}$ if $x_{k,b} \notin \mathcal{F}$ and as $x_{k,b}$ if $x_{k,b} \in \mathcal{F}$. Similarly we define $x_{k,c}$ as the intersection of the segment $[x_{k,b}, x_k^+]$ with the boundary of $\mathcal{F}$ if $x_{k,b} \notin \mathcal{F}$ and as $x_{k,b}$ if $x_{k,b} \in \mathcal{F}$. We may now verify (4.24) with the set $\{x_k, x_{k,a}, x_{k,c}, x_k^+\}$. Observe also that, if (4.24) fails, then there is a feasible local minimizer of the model along the path

$$\mathcal{P}_k \stackrel{\text{def}}{=} [x_k, x_{k,a}] \cup [x_{k,a}, x_{k,c}] \cup [x_{k,c}, x_k^+] \tag{4.26}$$

(the middle segment being possibly reduced to the point $x_{k,b}$ when it is feasible): further model minimization may then be started from this point in order to achieve AS5, ignoring the rest of the path and the trial point x_k^+. Note that $x_{k,b}$ is the solution of an essentially unconstrained model minimization (in the hyperplane orthogonal to g_k) and can be computed at reasonable cost, which makes checking this version of (4.24) acceptable from the computational point of view, especially since $x_{k,b}$ needs to be computed only once even if several x_k^+ must be tested. The definition of $x_{k,b}$ is not even necessary and other points $x_{k,b}$ are acceptable, as long as a suitable "descent path" $\mathcal{P}_k$ from x_k to x_k^+ can be determined. Figure 4.2 on the following page shows the path $\mathcal{P}_k$ given by (4.26) on a case where (4.5) fails. This figure also shows that there are cases where the only feasible model minimizer may be in a direction such that (4.5) fails.

Using AS6, we may now state the crucial lower bound on the model reduction.

Lemma 4.4 *Suppose that AS6 holds at iteration k. Then there exists a constant $\kappa_{\text{red}} > 0$ independent of k such that*

$$m_k(x_k) - m_k(x_k^+) \geq \kappa_{\text{red}} \sigma_k \|s_k\|^3. \tag{4.27}$$

Proof. If (4.5) holds, then the conclusion immediately follows from Lemma 4.2. Otherwise, we first note that

$$m_k(x_k) - m_k(x_{k,1}) \geq \tfrac{1}{6} \sigma_k \|x_{k,1} - x_k\|^3$$

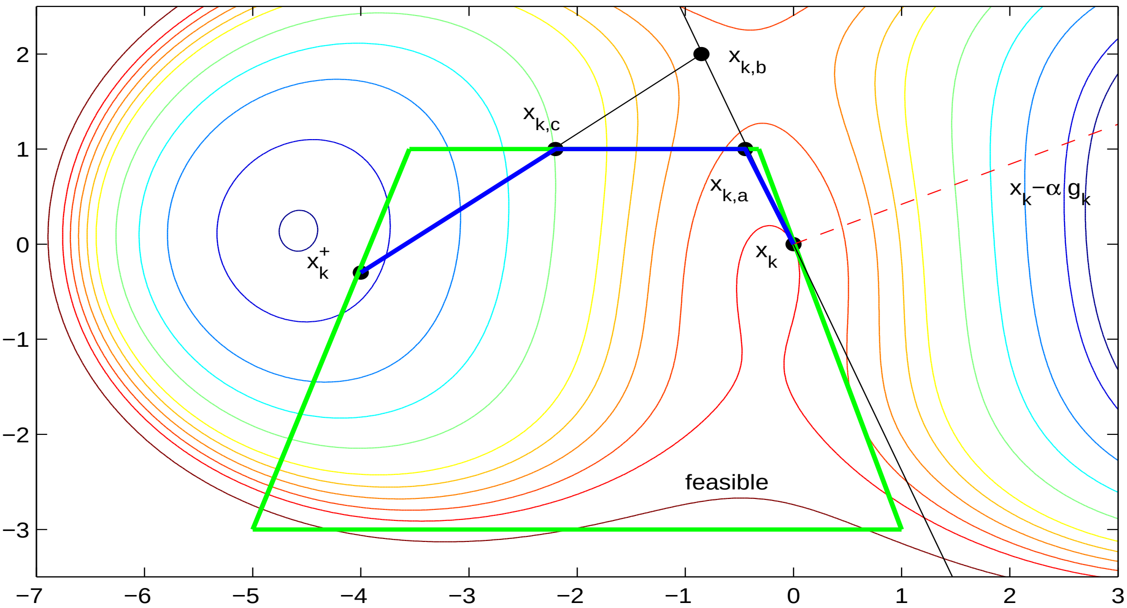

Figure 4.1: A constrained path $\mathcal{P}_k$ with $x_{k,a}$, $x_{k,b}$ and $x_{k,c}$ starting from the iterate $x_k = (0,0)^T$ on the cubic model $m(x,y) = -x - \frac{42}{100}y - \frac{3}{10}x^2 - \frac{1}{10}y^3 + \frac{1}{3}[x^2+y^2]^{\frac{3}{2}}$, the feasible set $\mathcal{F}$ being the polyhedron with vertices $(1,-5)^T$, $(-\frac{32}{100},1)^T$, $(-\frac{355}{100},1)^T$ and $(-\frac{510}{100},-5)^T$.

because of Lemma 4.2 and the fact that $x_{k,1}$ can be obtained by reducing the model along the segment $[x_k, x_{k,1}]$ implies the inequality $\langle g_k, x_{k,1} - x_k\rangle \le 0$. Moreover, AS6 implies that

$$m_k(x_{k,2}) \le m_k(x_{k,1}), \quad \langle \nabla_x m_k(x_{k,2}), x_{k,2} - x_{k,1}\rangle \le 0, \quad \text{and} \quad \langle \nabla_x m_k(x_{k,1}), x_{k,2} - x_{k,1}\rangle \le 0.$$

We may then apply Lemma 4.3 a first time with $x_\circ = x_{k,1}$ and $x_\bullet = x_{k,2}$ to deduce that

$$m_k(x_k) - m_k(x_{k,2}) \ge \tfrac{1}{6}\kappa_{\rm lm}\sigma_k\|x_{k,2} - x_k\|^3.$$

If $\ell_k > 2$, we then apply the same technique $\ell_k - 1$ times: for $i = 2,\ldots,\ell_k$, we deduce from AS6 that

$$m_k(x_{k,i}) \le m_k(x_{k,i-1}), \quad \langle \nabla_x m_k(x_{k,i}), x_{k,i} - x_{k,i-1}\rangle \le 0, \quad \text{and} \quad \langle \nabla_x m_k(x_{k,i-1}), x_{k,i} - x_{k,i-1}\rangle \le 0,$$

while we obtain by induction that

$$m_k(x_{k,i-1}) \le m_k(x_k) - \tfrac{1}{6}\kappa_{\rm lm}^{i-2}\sigma_k\|x_{k,i-1} - x_k\|^3.$$

This then allows us to apply Lemma 4.3 with $x_{k,\circ} = x_{k,i-1}$ and $x_{k,\bullet} = x_{k,i}$, yielding that

$$m_k(x_k) - m_k(x_{k,i}) \ge \tfrac{1}{6}\kappa_{\rm lm}^{i-1}\sigma_k\|x_{k,i} - x_k\|^3.$$

After $\ell_k - 1$ applications of Lemma 4.3, we obtain that

$$m_k(x_k) - m_k(x_{k,\ell_k}) \ge \tfrac{1}{6}\kappa_{\rm lm}^{\ell_k-1}\sigma_k\|x_{k,i} - x_k\|^3.$$

This is the desired bound with $\kappa_{\rm red} = \frac{1}{6}\kappa_{\rm lm}^{\bar{\ell}-1}$. □

We may then obtain an function-evaluation complexity result for Algorithm ACURC-S by completing our assumptions as follows.

AS7: The Hessian $H(x_k)$ is well approximated by B_k, in the sense that there exists a constant $\kappa_{\rm BH} > 0$ such that, for all k,

$$\|[B_k - H(x_k)]s_k\| \le \kappa_{\rm BH}\|s_k\|^2.$$

AS8: The Hessian of the objective function is "weakly" uniformly Lipschitz-continuous on the segments $[x_k, x_k + s_k]$, in the sense that there exists a constant $\kappa_{\rm LH} \geq 0$ such that, for all k and all $y \in [x_k, x_k + s_k]$,

$$\|[H(y) - H(x_k)]s_k\| \leq \kappa_{\rm LH}\|s_k\|^2.$$

AS9: The iterates of Algorithm ACURC-S remain in some bounded subset $\mathcal{F}_0 \subseteq \mathcal{F}$.

AS7 and AS8 are acceptable assumptions essentially corresponding to the cases analysed in Nesterov and Polyak (2006) and Cartis et al. (2007) for the unconstrained problem, the only differences being that the first authors assume $B_k = H(x_k)$ instead of the weaker AS7 and that AS8 is not expressed along the step s_k in the second reference. AS9 is only mildly restrictive, and is for instance satisfied if the feasible set $\mathcal{F}$ itself is bounded, or if the constrained level-set of the objective function $\{x \in \mathcal{F} | f(x) \leq f(x_0)\}$ is bounded. Note that AS9 implies AS3.

An important consequence of AS6-AS9 is that they to allow us to deduce the following crucial relation between local optimality and stepsize.

Lemma 4.5 *Suppose that AS1, AS2 and AS4-AS9 hold, that iteration k of Algorithm ACURC-S is successful and that*

$$\sigma_k \leq \sigma_{\max}, \tag{4.28}$$

for some constant $\sigma_{\max} > 0$ independent of k. Then, for some constant $\kappa_{\rm s} \in (0, 1)$ independent of k,

$$\|s_k\| \geq \kappa_{\rm s}\sqrt{\chi(x_k^+, 1)}. \tag{4.29}$$

Proof. We first consider the case where $x_k^+ = x_{k,1}^+$, again drop the index k for the proof, define $\chi^+ \stackrel{\rm def}{=} \chi(x_k^+, 1)$ and $g^+ \stackrel{\rm def}{=} g(x_k^+)$, and start by noting that

$$\begin{array}{rcl} \|g^+ - \nabla_x m(x_k^+)\| & = & \left\| g + \int_0^1 H(x + ts)s\,dt - g - [B - H(x)]s - H(x)s - \sigma\|s\|s \right\| \\ & = & \left\| \int_0^1 [H(x + ts) - H(x)]s\,dt \right\| + (\kappa_{\rm BH} + \sigma)\|s\|^2 \\ & \leq & \int_0^1 \|[H(x + ts) - H(x)]s\|\,dt + (\kappa_{\rm BH} + \sigma)\|s\|^2 \\ & \leq & (\kappa_{\rm LH} + \kappa_{\rm BH} + \sigma)\|s\|^2, \\ & \leq & (\kappa_{\rm LH} + \kappa_{\rm BH} + \sigma_{\max})\|s\|^2, \end{array} \tag{4.30}$$

where we have used (2.2), AS7, AS8, the triangular inequality and (4.28). Assume first that

$$\|s\| \geq \sqrt{\frac{\chi^+}{2(\kappa_{\rm LH} + \kappa_{\rm BH} + \sigma_{\max})}}. \tag{4.31}$$

In this case, (4.29) follows with $\kappa_{\rm s} = \sqrt{\frac{1}{2(\kappa_{\rm LH}+\kappa_{\rm BH}+\sigma_{\max})}}$, as desired. Assume therefore that (4.31) fails and observe that

$$\chi^+ \stackrel{\rm def}{=} |\langle g^+, d^+\rangle| \leq |\langle g^+ - \nabla_x m(x^+), d^+\rangle| + |\langle \nabla_x m(x^+), d^+\rangle| \tag{4.32}$$

where the first equality define the vector d^+ with

$$\|d^+\| \leq 1. \tag{4.33}$$

But, using the Cauchy-Schwartz inequality, (4.33), (4.30) the failure of (4.31) and (4.32) successively, we obtain that

$$\begin{array}{rcl} \langle \nabla_x m(x^+), d^+\rangle - \langle g^+, d^+\rangle & \leq & |\langle g^+, d^+\rangle - \langle \nabla_x m(x^+), d^+\rangle| \\ & \leq & \|g^+ - \nabla_x m(x^+)\| \\ & \leq & (\kappa_{\rm LH} + \kappa_{\rm BH} + \sigma_{\max})\|s\|^2 \\ & \leq & \frac{1}{2}\chi^+ \\ & = & -\frac{1}{2}\langle g^+, d^+\rangle, \end{array}$$

which in turn ensures that

$$\langle \nabla_x m(x^+), d^+ \rangle \le \tfrac{1}{2}\langle g^+, d^+ \rangle < 0.$$

Moreover, $x^+ + d^+ \in \mathcal{F}$ by definition of χ^+, and hence, using (4.33) and (4.23),

$$|\langle \nabla_x m(x^+), d^+ \rangle| \le \chi^{\rm m}(x^+, 1). \tag{4.34}$$

We may then substitute this bound in (4.32) and use the Cauchy-Schwartz inequality and (4.33) again to obtain that

$$\chi^+ \le \|g^+ - \nabla_x m(x^+)\| + \chi^{\rm m}(x^+, 1) \le \|g^+ - \nabla_x m(x^+)\| + \min(\kappa_{\rm stop}, \|s\|)\,\chi, \tag{4.35}$$

where the last inequality results from (4.22). We now observe that the successful nature of iteration k and AS9 imply that both x and x^+ belong to $\mathcal{F}_0$. Moreover, the inequality

$$\|g^+ - g\| = \left\| g + \int_0^1 H(x+ts)s\,dt - g \right\| \le \int_0^1 \|H(x+ts)\|\,\|s\|\,dt \le \kappa_{\rm H}\|s\|$$

(where we used the mean-value theorem, AS4 and the triangle inequality sucessively) and AS3, itself implied by AS9, yield that $\nabla_x f(x)$ is Lipschitz continuous on $\mathcal{F}_0$ with constant $\kappa_{{\rm L}g} = \kappa_{\rm H}$. Theorem 3.3 then ensures that $\chi(x)$ is Lipschitz continuous on $\mathcal{F}_0$ (with constant $\kappa_{{\rm L}\chi}$), and therefore that

$$\chi \le \kappa_{{\rm L}\chi}\|x - x^+\| + \chi^+ = \kappa_{{\rm L}\chi}\|s\| + \chi^+, \tag{4.36}$$

so that, using (4.35) and (4.30),

$$\chi^+ \le \|g^+ - \nabla_x m(x^+)\| + \kappa_{{\rm L}\chi}\|s\|^2 + \kappa_{\rm stop}\chi^+ \le (\kappa_{{\rm L}H} + \kappa_{\rm BH} + \sigma_{\max})\|s\|^2 + \kappa_{{\rm L}\chi}\|s\|^2 + \kappa_{\rm stop}\chi^+.$$

We thus deduce that

$$(1 - \kappa_{\rm stop})\chi^+ \le (\kappa_{{\rm L}H} + \kappa_{{\rm L}\chi} + \kappa_{\rm BH} + \sigma_{\max})\|s\|^2,$$

and therefore that

$$\|s\| \ge \sqrt{\frac{(1-\kappa_{\rm stop})\chi^+}{\kappa_{{\rm L}H} + \kappa_{{\rm L}\chi} + \kappa_{\rm BH} + \sigma_{\max}}}$$

which gives (4.29) with

$$\kappa_{\rm s} = \sqrt{\frac{1-\kappa_{\rm stop}}{\kappa_{{\rm L}H} + \kappa_{{\rm L}\chi} + \kappa_{\rm BH} + \sigma_{\max}}}. \tag{4.37}$$

□

We may now consolidate our result under our current assumptions.

Theorem 4.6 *Suppose that AS1, AS2 and AS4-AS9 hold, and that, for all k,*

$$\sigma_k \ge \sigma_{\min} \tag{4.38}$$

for some constant $\sigma_{\min} \in 0,1)$. *Then there exists a constant* $\kappa_{\rm df2} \in (0,1)$ *such that, for every* $k \ge 0$, $k \in \mathcal{S}$,

$$f(x_k) - m_k(x_k^+) \ge \kappa_{\rm df2}\chi_{k+1}^{\frac{3}{2}}. \tag{4.39}$$

As consequence, the ACURC-S algorithm needs at most

$$\left\lceil \kappa_{\mathcal{S}}\epsilon^{-\frac{3}{2}} \right\rceil$$

successful iterations and evaluations of $\nabla_x f$ *and (possibly)* $\nabla_{xx} f$, *and at most*

$$\left\lceil \kappa_{*}\epsilon^{-\frac{3}{2}} \right\rceil$$

iterations and objective function evaluations to terminate, that is to generate an iterate x_k *such that* $\chi_k \le \epsilon \le 1$, *where*

$$\kappa_{\mathcal{S}} \stackrel{\rm def}{=} \left\lceil \frac{f(x_0) - f_{\rm low}}{\eta_1 \kappa_{\rm df}} \right\rceil \quad \textit{and} \quad \kappa_* \stackrel{\rm def}{=} \kappa_{\mathcal{S}} + (1+\kappa_{\mathcal{S}})\frac{\log\left(\max\left[\sigma_0, \frac{3\gamma_2\kappa_{{\rm L}H}}{2}\right]/\sigma_{\min}\right)}{\log(\gamma_1)}.$$

Proof. We first recall that the mechanism of the algorithm ensures that (4.5) holds for each step s_k, and thus, by Lemma 4.2, that (4.6) holds for all k. We then deduce from Lemma 5.2 in Cartis et al. (2007), itself strongly relying on AS7 and AS8, that

$$\sigma_k \leq \max\left[\sigma_0, \frac{3\gamma_2\kappa_{\rm LH}}{2}\right] \stackrel{\rm def}{=} \sigma_{\max}.$$

This allows us to apply Lemma 4.5 with this upper bound on σ_k. We then obtain from (4.27) and (4.29) that

$$f(x_k) - m_k(x_k^+) \geq \tfrac{1}{6}\sigma_{\min}\kappa_{\rm red}\kappa_{\rm s}^3\chi_{k+1}^{\frac{3}{2}},$$

which is (4.39) with $\kappa_{\rm df} \stackrel{\rm def}{=} \sigma_{\min}\kappa_{\rm red}\kappa_{\rm s}^3$, where $\kappa_{\rm s}$ is given by (4.37). The second conclusion of the theorem then follows from Theorems 6.1 and 6.2 in Cartis et al. (2007). □

This result shows a worst-case complexity result in terms of evaluations of the problem's functions which is of the same order as that for the unconstrained case (see Nesterov and Polyak, 2006, or Cartis et al., 2007).

We conclude our analysis by observing that global convergence to first-order critical points may be ensured for Algorithm ACURC-S (even without AS5-AS8), if one simply ensure that the steps s_k ensures a model decrease which is larger than that obtained at the Cauchy point (as computed by Step 1 of Algorithm ACURC), which means that (2.9) must hold, a very acceptable condition. The convergence analysis presented for Algorithm ACURC thus applies without modification.

4.3 Solving the subproblem

For the better complexity bound of Theorem 4.6 to hold, we need, on each iteration k, to approximately and iteratively minimize the model $m_k(s)$ in $\mathcal{F}$ along a uniformly bounded number of line segments so as to ensure AS6, until condition (4.22) is satisfied. Active-set techniques may be applied to $m_k(x)$, starting at x_k, a minimal and simple such approach being the basic ACURC$_\epsilon$ framework (applied to m_k). Though in practice, a (much) more efficient active-set technique should be employed, its theoretical guarantees of finite termination for such methods seems nontrivial to derive in the context of AS5 and AS6, due to the combinatorial aspect of both the (nonconvex) objective and the constraints. Thus for now, let us briefly discuss in more detail applying ACURC$_\epsilon$ to m_k starting at x_k, for each $k \geq 0$. Let us assume in what follows that $k \geq 0$ is fixed. In particular, note that we terminate each application of ACURC$_\epsilon$ to m_k when AS5 is satisfied. As the latter depends on χ_k, it is appropriate that we deduce a lower bound on $m_k(x)$, $x \in \mathcal{F}$ that also depends on χ_k.

Lemma 4.7 *Let AS1 – AS2, AS4 and (4.38) hold. Then*

$$m_k(x_k + s) - f(x_k) \geq -\kappa_{\rm lm}\max[\kappa_{\rm B}^2, \kappa_{\rm B}\chi_k, (\chi_k)^{3/2}], \quad x_k + s \in \mathcal{F}, \quad k \geq 0, \tag{4.40}$$

where $\kappa_{\rm lm} = 18\kappa_{\rm B}/\sigma_{\min}^2$.

Proof. Letting $x_k + s \in \mathcal{F}$, we have from (3.4) that

$$\langle g_k, s\rangle \geq -\chi(x_k, \|s\|) \geq -\chi_k \max[\|s\|, 1],$$

where in the second inequality, we used (3.9) and (3.10). It follows from (2.2) and AS4 that

$$\begin{aligned}
m_k(x_k + s) - f(x_k) &= \langle g_k, s\rangle + \tfrac{1}{2}\langle s, B_k s\rangle + \tfrac{1}{3}\sigma_k\|s\|^3 \\
&\geq -\chi_k \max\left[\|s\|, 1\right] - \kappa_{\rm B}\|s\|^2 \\
&\geq -\max\left[\|s\|, 1\right](\chi_k + \kappa_{\rm B}\|s\|) \\
&\geq -2\max\left[\|s\|, 1\right] \cdot \max\left[\chi_k, \kappa_{\rm B}\|s\|\right] \\
&\geq -2\kappa_{\rm B}\max\left[\|s\|, 1\right] \cdot \max[\chi_k, \|s\|],
\end{aligned} \tag{4.41}$$

where in the last inequality, we employed $\kappa_{\rm B} > 1$. Note that it is, in fact, sufficient to consider points for which $m_k(x_k+s) \le f(x_k)$, as for the others, m_k is bounded below by $f(x_k)$. This and an argument similar to that of Lemma 3.6 yield

$$\|s\| \le \frac{3}{\sigma_k} \max\left[\kappa_{\rm B}, (\sigma_k \chi_k)^{1/2}, (\sigma_k\chi_k)^{1/3}\right],$$

and furthermore, from (4.38), $\sigma_{\min} \in (0,1)$ and $\kappa_{\rm B} > 1$, we obtain that

$$\|s\| \le \frac{3}{\sigma_{\min}} \max\left[\kappa_{\rm B}, \chi_k^{1/2}, \chi_k^{1/3}\right], \quad \max\left[\|s\|, 1\right] \le \frac{3}{\sigma_{\min}} \max\left[\kappa_{\rm B}, \chi_k, \chi_k^{1/2}, \chi_k^{1/3}\right].$$

Substituting the above bounds into the last inequality in (4.41) yields (4.40). □

When applying ACURC$_\epsilon$ to m_k, we need to iterate until (4.22) holds, namely the tolerance for the first-order optimality measure is set to

$$\epsilon_k := \min\{\kappa_{\rm stop}, \|s_k\|\}\chi_k. \tag{4.42}$$

In order to estimate the complexity of employing ACURC$_\epsilon$ to m_k with the above tolerance, we apply Theorem 4.1 with $f := m_k$ and $\epsilon := \epsilon_k$. Furthermore, the gap $f(x_0) - f_{\rm low}$ is now $f(x_k) - m_{k,{\rm low}}$, for which (4.40) gives an upper bound. Note that from (4.42), (4.40) and Theorem 4.1, if the stepsize s_k or χ_k are large, then the complexity bound is of order χ_k^{-2} or better.

To better quantify this bound on the iteration count, recall that from (4.29), for successful k, we have

$$\epsilon \ge \min\{\kappa_{\rm stop}, \kappa_{\rm s}\sqrt{\chi_{k+1}}\}\chi_k \ge \kappa_0 \min\{1, \sqrt{\omega_k}\}\omega_k \stackrel{\rm def}{=} \underline{\epsilon}_k,$$

where $\omega_k := \min\{\chi_{k+1}, \chi_k\}$ and $\kappa_0 > 0$. Thus if $\chi_k^m(x_{k+1}) \le \underline{\epsilon}_k$, then AS5 holds. Now we can use $\underline{\epsilon}_k$ in place of ϵ in Theorem 4.1, and deduce order $\underline{\epsilon}_k^{-2}$ inner-iteration worst-case complexity bound.

Note that applying ACURC$_\epsilon$ implies constructing local cubic models for m_k. However, m_k has a Lipschitz continuous Hessian with Lipschitz constant $(1+\sqrt{2})\sigma_k$, as we now show.

Lemma 4.8 *Consider the cubic model $m_k(x_k+s)$, $s \in \mathbb{R}^n$, in (2.2) for any fixed $k \ge 0$. Then the Hessian $\nabla_{xx} m_k(x_k+s)$ is globally Lipschitz continuous with Lipschitz constant $(1+\sqrt{2})\sigma_k$, namely*

$$\|\nabla_{xx} m_k(x_k+s) - \nabla_{xx} m_k(x_k+y)\| \le (1+\sqrt{2})\sigma_k\|s-y\|, \quad \forall s, y \in \mathbb{R}^n. \tag{4.43}$$

Proof. From (2.2), we have that

$$\nabla_{xx} m_k(x_k+s) = B_k + \sigma_k\|s\|I + \sigma_k\frac{ss^T}{\|s\|}.$$

Let $s, y \in \mathbb{R}^n$. Then

$$\begin{aligned}
\|\nabla_{xx} m_k(x_k+s) - \nabla_{xx} m_k(x_k+y)\| &= \left\|\sigma_k(\|s\|-\|y\|)I + \sigma_k\left(\tfrac{ss^T}{\|s\|} - \tfrac{yy^T}{\|y\|}\right)\right\| \\
&\le \sigma_k|\|s\|-\|y\|| + \sigma_k\left\|\|s\|\left(\tfrac{s}{\|s\|}\right)\left(\tfrac{s}{\|s\|}\right)^T - \|y\|\left(\tfrac{y}{\|y\|}\right)\left(\tfrac{y}{\|y\|}\right)^T\right\| \\
&\le \sigma_k\|s-y\| + \sigma_k\left\|\|s\|uu^T - \|y\|ww^T\right\|,
\end{aligned}$$

where $u \stackrel{\rm def}{=} s/\|s\|$ and $w \stackrel{\rm def}{=} y/\|y\|$. Thus (4.43) follows provided we show that

$$\left\|\|s\|uu^T - \|y\|ww^T\right\| \le \sqrt{2}\|s-y\|. \tag{4.44}$$

Letting $A \stackrel{\rm def}{=} \|s\|uu^T - \|y\|ww^T$, we have that

$$\begin{aligned}
A^TA &= \|s\|^2uu^T - \|s\|\cdot\|y\|u^Tw[uw^T + wu^T] + \|y\|^2ww^T \\
&= ss^T - \tfrac{\langle s,y\rangle}{\|s\|\cdot\|y\|}[sy^T + ys^T] + yy^T \\
&= ss^T - [sy^T + ys^T] + yy^T + \left(1 - \tfrac{\langle s,y\rangle}{\|s\|\cdot\|y\|}\right)[sy^T + ys^T] \\
&= (s-y)(s-y)^T + \left(1 - \tfrac{\langle s,y\rangle}{\|s\|\cdot\|y\|}\right)[sy^T + ys^T].
\end{aligned}$$

Thus, using Cauchy-Schwarz inequality to ensure $1 - \langle s, y\rangle/(\|s\| \cdot \|y\|) \geq 0$, we have that

$$\begin{aligned} \|A^T A\| &\leq \|s-y\|^2 + 2\|s\| \cdot \|y\| \left(1 - \tfrac{\langle s,y\rangle}{\|s\|\cdot\|y\|}\right) \\ &= \|s-y\|^2 + 2\left(\|s\| \cdot \|y\| - \langle s, y\rangle\right). \end{aligned}$$

But $2\,[\,\|s\| \cdot \|y\| - \langle s, y\rangle\,] \leq \|s-y\|^2$, and so (4.44) follows by using that A is symmetric and hence $\|A\|^2 = \|A^T A\|$. □

As a consequence of this observation, we may keep the "low-level" values of the cubic regularisation parameters fixed at some multiple of σ_k larger than $1+\sqrt{2}$ and then all iterations of ACURC$_\epsilon$ applied to m_k are very successful. Furthermore, the upper bound (3.52) on the "low-level" cubic parameters is now independent on the accuracy tolerance of the subproblem.

The iteration complexity of solving the subproblem may seem discouraging at first sight, but one has to remember that we have used a very naive algorithm for this purpose, and it does not involve the problem's nonlinear objective function at all.

5 Conclusions and perspectives

We have generalized the adaptive cubic overestimation method for unconstrained optimization to the case where convex constraints are present. Our method is based on the use of the orthogonal projector onto the feasible domain, and is therefore practically limited to situations where applying this projector is computationally inexpensive. This is for instance the case if the constraints are simple lower and upper bounds on the variables, or if the feasible domain has a special shape such as a sphere, a cylinder or the order simplex (see Section 12.1.2 of Conn et al., 2000). The resulting ACURC algorithm has been proved globally convergent to first-order critical points. This result has capitalized on the natural definition of the first-order criticality measure (3.1), which allows a reasonably easy extension of the unconstrained proof techniques to the constrained case. As a by-product, the Lipschitz continuity of the criticality measure $\chi(x)$ has also been proved for bounded convex feasible sets.

A variant of Algorithm ACURC has then been presented for which a worst-case function-evaluation complexity bound can be shown, which is of the same order as that known for the unconstrained case. Remarkably, this algorithm does not rely on global model minimization, but the result obtained is only in terms of the global number of iterations and problem's function's evaluations, leaving aside the complexity of solving the subproblem, even approximately.

The authors are well aware that many issues remain open at this stage, amongst which the details of an effective step computation, the convergence to second-order points and the associated rate of convergence and the constraint identification properties, as well as the implications of the new complexity result on optimization with equality and inequality constraints. Numerical experience is also necessary to assess the practical potential of both algorithms.

Acknowledgements

The authors are grateful to Ken McKinnon for supplying the main idea for the proof of Lemma 3.2, and to Dimitri Tomanos for carefully reading an early draft of this paper. The work of the third author was also supported in part by the EPSRC grants EP/E053351/1, EP/F005369/1 and EP/G038643/1 and by the European Science Foundation through the OPTPDE program.

References

C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Technical Report 07/05, Department of Mathematics, FUNDP - University of Namur, Namur, Belgium, 2007.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization with simple bounds. *SIAM Journal on Numerical Analysis*, **25**(182), 433–460, 1988. See also same journal 26:764–767, 1989.

A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 01 *in* 'MPS-SIAM Series on Optimization'. SIAM, Philadelphia, USA, 2000.

A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, **3**(1), 164–221, 1993.

S. Gratton, M. Mouffe, Ph. L. Toint, and M. Weber-Mendonça. A recursive trust-region method in infinity norm for bound-constrained nonlinear optimization. *IMA Journal of Numerical Analysis*, **28**(4), 827–861, 2008*a*.

S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, **19**(1), 414–444, 2008*b*.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, 1981.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms. Part 1: Fundamentals*. Springer Verlag, Heidelberg, Berlin, New York, 1993.

O. L. Mangasarian and J. B. Rosen. Inequalities for stochastic nonlinear programming problems. *Operations Research*, **12**(1), 143–154, 1964.

Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Applied Optimization. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Yu. Nesterov. Cubic regularization of Newton's method for convex problems with constraints. Technical Report 2006/9, CORE, UCL, Louvain-la-Neuve, Belgium, 2006.

Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, **108**(1), 177–205, 2006.

R. T. Rockafellar. *Convex Analyis*. Princeton University Press, Princeton, USA, 1970.

A. Sartenaer. Armijo-type condition for the determination of a generalized Cauchy point in trust region algorithms using exact or inexact projections on convex constraints. *Belgian Journal of Operations Research, Statistics and Computer Science*, **33**(4), 61–75, 1993.

M. Weiser, P. Deuflhard, and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, **22**(3), 413–431, 2007.